Essays in Industrial Organization and Econometrics

Essays in Industrial Organization and Econometrics

A dissertation presented

by

Fanyin Zheng

to

The Department of Economics

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Economics

Harvard University

Cambridge, Massachusetts

May 2015

Dissertation Advisor:
Professor Ariel Pakes

Author:
Fanyin Zheng

Essays in Industrial Organization and Econometrics

Abstract

This dissertation consists of three essays, two on estimating dynamic entry games and one on the inference for misspecified models with fixed regressors.

Big box retail stores have large impact on local economies and receive large subsidies from local governments. Hence it is important to understand how discount retail chains choose store locations. In the first two essays, I study the entry decisions of those firms, examine the role of preemptive incentives, and evaluate the impact of government subsidies on those decisions. To quantify preemptive incentives, I model firms' entry decisions using a dynamic duopoly location game. Stores compete over the shopping-dollars of close-by consumers, making store profitability spatially interdependent. I use separability and two-stage budgeting to reduce the state space of the game and make the model tractable. Instead of adopting census geographic units, I infer market divisions from data using a clustering algorithm built on separability conditions. I introduce a 'rolling window' approximation to compute the value function and estimate the parameters of the game. The results suggest that preemptive incentives are important in chain stores' location decisions and that they lead to loss of production efficiency. On average, the combined sum of current and future profits of the two firms is lowered by 1 million dollars per store. Finally, I assess the impact of government subsidies to encourage entry when one retailer exits, as happened in the recent crisis. I find that although the welfare loss such exits cause on local economies can be substantial, the average size of observed subsidies is not enough to affect firms' entry decisions. This study is organized as follows. In the first essay, I provide descriptive evidence of preemptive entry in the discount retail industry and explain how I model firms'

entry decisions. In the second essay, I describe the estimation strategy and present the counterfactual analyses.

The third essay is joint work with Alberto Abadie and Guido Imbens. Following the work by Eicker (1967), Huber (1967) and White (1980ab; 1982) it is common in empirical work to report standard errors that are robust against general misspecification. In a regression setting these standard errors are valid for the parameter that minimizes the squared difference between the conditional expectation and the linear approximation, averaged over the population distribution of the covariates. In this essay, we discuss an alternative parameter that corresponds to the approximation to the conditional expectation based on minimization of the squared difference averaged over the sample, rather than the population, distribution of the covariates. We argue that in some cases this may be a more interesting parameter. We derive the asymptotic variance for this parameter, which is generally smaller than the Eicker-Huber-White robust variance, and propose a consistent estimator for this asymptotic variance.

Contents

List of Tables

List of Figures

Acknowledgments

I am deeply grateful for the guidance and encouragement I received from Ariel Pakes, Guido Imbens, and Greg Lewis. Their advice and support throughout my graduate school have been invaluable. Working with them has been a truly intellectually stimulating and rewarding experience.

I would like to thank Elie Tamer for many helpful conversations and suggestions. I am grateful to Alberto Abadie, Jean Baccelli, Gary Chamberlain, Max Kasy, Zhenyu Lai, Danial Lashkari, Robin Lee, Jing Li, Hong Luo, Eric Maskin, Eduardo Morales, Daniel Pollmann, Marc Rysman, Mark Shepard, Che-Lin Su, Tom Wollmann, Jing Xia, Lilei Xu, Ali Yurukoglu, Tom Zimmermann, and participants in the industrial organization and econometrics seminar at Harvard University for helpful comments and suggestions on the content in the first two chapters of this dissertation. I would like to thank Guido Imbens and Alberto Abadie for the coauthorship of the third chapter of this dissertation. I am also grateful for comments by Hal White and participants in the econometrics lunch seminar at Harvard University, and in particular for discussions with Gary Chamberlain on the content in this chapter. Financial support from the Department of Economics and Lab for Economic Applications and Policy at Harvard University, the Douglas Dillon Fellowship, and the Chiles Foundation Scholarship is gratefully acknowledged.

I would like to thank Jim Walker, Jack Porter, and Larry Samuelson for introducing me to economic research. To my friends and colleagues in graduate school, I am grateful for their generous support and friendship in the past six years. The journey would not have been the same without them. I am also grateful to Jean Baccelli for his patience, love, and support through the process of writing this dissertation.

Finally I would like to thank my parents, Feng Li and Zheng Gang, for the unwavering love, inspiration, and encouragement.

To my parents

Chapter 1

Spatial Competition and Preemptive Entry in the Discount Retail Industry: Descriptive Evidence and Model

1.1 Introduction

The discount retail industry has been a fast growing sector of the U.S. economy since the 1960s. Back in 1962, only three small chains existed with less than 200 stores in total. Today, there are many more national chains with over 5000 stores in the country, generating revenue of over a hundred billion dollars per year. Such fast growth has had a large impact on local economies. On the one hand, consumers benefit from the low prices and product varieties of stores such as Walmart and Kmart. New stores of discount retailers also boost local employment. However, small businesses and other retailers suffer from the presence of discount retailers (Jia, 2008). Overall local employment rates may be lower due to entry by a discount retailer (Basker, 2007; Neumark *et al.*, 2008). Employees criticize firms such as Walmart for driving down wages and benefits (Basker, 2007). In small towns, residents complain about dying main streets and business centers. These studies show that discount retailer's entry has a large impact on local economies.

Multi-store retail chains are also receiving large amounts of subsidies from local governments in the form of sales tax rebate, property tax rebate, infrastructure assistance, etc.. Walmart alone received over 160 million dollars in the past 15 years[1]. Local economies are affected as much by store closings as by store openings, as shown by Kmart closing over 1000 stores in the wake of the recent crisis[2]. Local governments are proposing to subsidize other retailers to replace closing stores. However, most of the abandoned retail space, such as the former Kmart stores, has remained empty for years[3]. Whether subsidies affect discount retailers' entry decisions and more generally how the firms make entry decisions thus become an important question for policy makers. The first goal of this study is to study how multi-store retail chains such as Walmart and Kmart make entry decisions.

To answer this question, I use a data set about two discount retailers. Both retailers are among the largest in the country. Since part of the data is proprietary, I am not able to reveal the identities of the firms. In what follows, I will refer to them as Blue firm and Red firm. The two firms were among the first to open discount stores in the U.S. and both experienced long periods of fast growth. Blue firm was much smaller than Red firm before the 1980s, but surpassed Red firm in the early 1990s and became one of the largest employers in the country. Blue firm succeeded in competing against Red firm because it carefully chose store locations, exploited economies of density and, most interestingly, possibly made preemptive entry moves (Bradley *et al.*, 2002; Holmes, 2011). That is, Blue firm might have entered earlier in markets in which it feared Red firm would enter. As a consequence, the second goal of this study is to investigate preemptive incentives and to quantify their impact on multi-store retailers' entry decisions and on the production efficiency of opening new stores. The definition of preemptive entry I will follow hinges on how much (in equilibrium) the likelihood of one firm entering a particular location today is impacted by the likelihood of its opponent entering the same location in the future, holding its static profits constant.

[1]Source: goodjobfirst.org.

[2]Source: Kmart annual reports.

[3]Source: `http://www.floridatoday.com/story/money/business/2014/07/27/kmart-goes-next/13197001/`

The data I use consists of two major parts. The first part contains store level sales data and consumer demographic data, which will be used in the demand estimation. The second part is firms' entry data which contains geocoded store locations and store opening dates of each Blue and Red store opened between 1985 and 2001. Blue firm's entry data comes from Holmes (2011). I collected Red firm's entry data using various sources including yellow page data and industry journals.

The model through which I analyze my data has two main features. First, it allows strategic interactions between firms in a dynamic duopoly framework. This is necessary because preemptive incentives cannot be studied in either a dynamic single-agent or a static game setting. Second, stores are spatially interdependent through demand and the firm-level entry decisions. This last feature fits the nature of the discount retail industry, since firms operate multiple stores that often locate close to each other. It also applies to other retail industries or markets in which firms operate in multiple interdependent locations or sectors.

As in many empirical models of dynamic games, a major obstacle to estimation is computing the value function for a large number of possible choice paths. The problem is particularly difficult to solve in the current setting given that entry decisions are made at the firm level and that decisions are not independent across markets. Therefore, I develop a series of tools to make the model tractable.

First, I apply two-stage budgeting and separability conditions to decentralize firms' entry decisions across markets. Conditional on optimal market-level budget, if markets are separable, entry decisions become optimal within each market. This allows me to condition on the observed budget constraint of each market and solve the game for each market independently. Then, I build a clustering algorithm based on the separability conditions and apply it to partition the national market, while preserving the spatial interdependence across stores within each market. Finally, I employ a 'rolling window' approximation to compute value functions. That is, instead of optimizing over an infinite horizon, firms optimize over a fixed number of periods ahead and approximate the continuation value

using scaled terminal values. The set of potential paths of choices each firm is optimizing over is therefore restricted, but the approximation is consistent with how managers actually make decisions. Due to the non-stationary nature of the problem at hand, the dynamic game cannot be estimated in a two-stage procedure as in (Bajari *et al.*, 2007) (BBL) or Pakes *et al.* (2007) (POB). Accordingly, I solve for the nested fixed point in the estimation as in Pakes (1986) and Rust (1987). The parameter estimates are obtained by solving the game using backwards induction and maximizing the likelihood of observed location choices in each market and each period.

Using the estimated parameters, I conduct counterfactual analyses to quantify preemptive incentives and to evaluate subsidy policies. The first counterfactual analysis quantifies preemptive incentives by removing them from one firm's optimization problem and comparing it to the original equilibrium. The challenge is that preemption is a motive instead of an action and thus it is difficult to be distinguished from other optimization motives in the entry decision. Accordingly, I use a one-period deviation approach to identify preemption. The preemptive motives of Blue firm are removed from the optimization motives of its entry decision by taking Blue firm's observed choices out of Red firm's choice set for one period. The reasoning is as follows: if Blue firm chose the observed locations in the current period because it feared Red firm would enter otherwise, Blue firm should delay entry at those locations now, since Red firm is not allowed to enter in the following period. Results show that preemption costs Blue firm 0.86 million dollars per store on average, which is equivalent to a small store's one year profits. The combined current and future profits of the two firms increase by 397 million dollars when preemption is removed, which is about 1 million dollars per store. The findings thus suggest that preemptive incentives are important to multi-store retailers' entry decisions and that preemptive entry can lead to substantial production efficiency loss.

In a second counterfactual analysis, I evaluate the subsidy policies proposed by local governments to encourage entry by Blue firm during a period in which Red firm exited many markets. I find that the average level of subsidies is not enough to induce entry and

that preemptive incentives affect the level of subsidies Blue firm needs to enter. Finally, I compute consumer welfare loss from longer travel time to shops when a Red store closes. I find that the welfare loss can be as big as the average size of the observed subsidies Blue firm received in the past.

This study contributes to a literature studying the discount retail industry. Holmes (2011) showed the importance of economies of scale in Walmart's expansion, using a single-agent dynamic optimization model. Jia (2008) studied the impact of Walmart and Kmart on small business, by solving a static game between Walmart and Kmart. Ellickson *et al.* (2013) and Zhu and Singh (2009) also studied economies of scale and competition between big chains, in a static setting. This study complements the literature by presenting a dynamic duopoly model to investigate the dynamic strategic interactions between firms while preserving features such as economies of scale and spatial competition that the papers mentioned above studied. In addition, the modeling and estimating methods in this paper make it possible to conduct counterfactual analyses to quantify preemptive incentives and evaluate subsidy policies.

The entry literature has been pioneered by Bresnahan and Reiss (1991) and Berry (1992). In most of the literature, for example in Mazzeo (2002) and in Seim (2006), firms make independent entry decisions in each market. In this paper by contrast, entry decisions are made at the firm level and markets are spatially interdependent. Jia (2008) allows interdependence across entry decisions, but the interdependence is assumed to be positive and linear in store density. More general forms of interdependence are allowed in this paper. They are also explicitly modeled through demand and firm level budget constraints.

This study also contributes to a recent empirical literature on preemptive incentives. Schmidt-Dengler (2006) studied preemptive incentives in the adoption of MRI by hospitals. He identifies preemptive incentives by solving a pre-commitment game and comparing the result to the original equilibrium in which players are allowed to respond to the opponent's action in each period. Igami and Yang (2014) examine burger chains' preemptive entry decisions, by solving a single agent's dynamic optimization problem and comparing the

results to the dynamic duopoly equilibrium. By contrast, this study introduces a one-period deviation method to identify preemptive incentives, while allowing static strategic interactions between firms and keeping payoffs comparable.

As for the theoretical tools used in the study, the two-stage budgeting and separability results come from classic theorems by Gorman (1971) on consumption problems. This study generalizes the main theorems in Gorman (1959, 1971), so that they can be applied in a dynamic game setting. The clustering algorithm developed in the study is based on those separability results. It belongs to the class of greedy algorithms of the graph partitioning literature (Fortunato and Castellano, 2012). It is applicable to other graph partitioning or market division problems in which geographic contiguity is preserved and precision of the solution is preferred to speed.

The study is organized as follows. Section 1.2 introduces the background of the industry in more detail, describes the data and provides descriptive evidence of preemptive incentives. Section 1.3 introduces the model, the application of two-stage budgeting and separability, and explains how markets can be defined using machine learning tools. Chapter 2 presents the empirical part of the study including the counterfactual analyses. Section 2.2 shows how the value functions can be approximated and presents the estimation results. Counterfactuals under which preemptive motives are removed are presented in Section 2.3. The subsidy policy application is presented in Section 2.4. Section 2.5 concludes.

1.2 Industry Background and Data

1.2.1 Discount Retail Industry: Background

The discount retail industry in the U.S. started when Walmart and Kmart opened their first stores in 1962. It has been growing very fast in the following 40 years. The total sales of discount stores peaked at 137 billion dollars in 2001 (Census, Annual Retail Trade Survey). The discount retail industry is a very concentrated one. In 2002, the four largest firms controlled 95% of sales (Census, Economic Census).

Table 1.1: *Comparison of Blue firm and Red firm in 2001*

	Blue	Red
Number of stores	2698	1883
Number of distribution centers	35	18

The two firms this paper studies, Blue and Red firm, are among those four largest firms. They followed the path of growth of the industry. Blue firm has had a particularly interesting pattern of growth. It was very small at the beginning of the industry, with less than 300 stores in the 1980s when Red firm already had over 1000 stores. But it surpassed Red firm in the 1990s and became one of the largest employers in the country. Table 1.1 presents the total number of stores and distribution centers of Blue and Red firm in 2001. Blue firm appears to be much bigger than Red firm in both dimensions. To explain Blue firm's success, researchers have highlighted carefully chosen store locations, efficient distribution network, high store density, and economies of scale (Bradley *et al.*, 2002; Holmes, 2011). Since Blue and Red firm compete in the same market, it is natural to examine whether these characteristics have also played a role in Blue firm's surpassing Red firm. In his study of economies of scale, Holmes (2011) raises the additional question of possible preemptive entry - a topic this paper will be concerned with.

Discount retail stores are known to have a large impact on local economies. Consumers benefit from the low prices of discount stores. Ellickson and Misra (2008) find that when Walmart enters a market, its low prices extend to other local stores. Basker (2007) shows that local employment is boosted after Walmart's entry. The impact is not always positive, however. Jia (2008) finds that half of the decline of small businesses in U.S. is caused by entry of Walmart or Kmart during the 1980s and 1990s. Basker (2007) also shows that when Walmart opens a new store, local employment shrinks in the long term due to the closings of small businesses. Because of the large and complex impact of discount retailers on the local economy, it is in the interest of policy makers to understand how decisions about

where to locate stores are made - the issue this paper investigates.

Discount retailers also turn out to receive large amounts of subsidies from local governments. According to goodjobsfirst.org, Walmart alone received over 160 million dollars between 2000 and 2014. The subsidies take on various forms including sales tax rebate, property tax rebate, free land, infrastructure assistance, etc.. Since Red firm started exiting many markets in 2001, local governments have been proposing subsidies to Red firm so that it would stay or to other retailers like Blue firm so that they would enter. For example, Buffalo, NY, proposed a 400,000 dollar subsidy to Red firm for it to stays[4]. Lots of retail space stayed empty for years. In Rockledge, FL, for example, the ex-Red store has been empty for 11 years[5]. It is not clear if the proposed size of subsidies is big enough to affect retailers' entry decisions in general - a question this paper will assess.

1.2.2 Data

Data limitations of the discount retail industry heavily constrains the models that can be used to analyze it. This is why I describe the data sources before presenting the model.

There are four main components of the data. The first component is store and distribution center locations and time of opening between 1985 and 2003. Blue firm's store and distribution center locations and opening dates between 1985 and 2003 come from Holmes (2011). Red firm's locations and time of opening data come from three sources. First, addresses and time of store openings are from infoUSA in 2002[6]. Second, I double-checked the addresses and time of opening of each store using the annual Chain Store Guide between 1984 and 2001. This step was necessary because there are 96 Red closings after 2000, and some of the stores are missing in the 2002 InfoUSA data. I do not model store closing decisions in this paper, but in the policy application in Section 2.4, I will discuss entry after

[4]Source: `www.huffingtonpost.com/2012/01/26/sears-closes-cities_n_1231326.html`

[5]Source: `www.floridatoday.com/story/money/business/2014/07/27/kmart-goes-next/13197001/`

[6]Red firm stopped opening stores after 2002.

a store closure. The time of opening and closing of theses missing stores was collected by searching through local newspapers[7]. Finally, I geocoded store addresses using the ArcGIS North America Address Locator. The distribution center addresses of Red firm have been collected from data published by the U.S. Environmental Protection Agency (EPA)[8]. The addresses have also been geocoded using ArcGIS, and opening dates have been collected from local newspapers.

Figure 1.1 presents the store and distribution centers of Blue firm on the map of contiguous U.S., as a snapshot by the end of 2001. The blue dots indicate Blue stores and the green diamonds indicate Blue distribution centers. Figure 1.2 presents the stores and distribution centers of Red firm by 2001. Each red dot is a Red store and each yellow diamond is a Red distribution center. Comparing the two maps, it appears that Blue firm has both more stores and more distribution centers, while Red stores seem to be more concentrated geographically. The figures also show that both firms are national chains and they compete in many local markets across the nation.

The sample consists of Blue and Red store openings between 1995 and 2001. 1984 and 1140 Blue and Red stores opened in this period, respectively. Store openings between 2002 and 2003 are left out of the sample because Red firm stopped opening new stores in 2002. Figures 1.3 and 1.4 display the sample store openings by year. It appears that Blue firm opened more stores than Red firm in almost every year[9].

Table 1.2 provides the summary statistics of the characteristics of the sample by firm. The characteristics are measured for the median store in 2001. First of all, it appears that the median distance to the closest competitor's store for Blue stores, 8.38 miles, is much bigger than for Red ones, 3.46 miles. The difference suggests that Red stores face more competition from Blue stores than Blue ones do from Red ones. Comparing this difference to the smaller

[7]For the 12 of Red stores that I could not find information about, I assumed the time of opening to be the first quarter of the year it first appeared in Chain Store Guide, and the time of closing to be the first quarter of the year they first disappeared.

[8]Distribution centers are EPA regulated facilities.

[9]The peak for Red firm in 1992 corresponds to the acquisition of a small chain. The stores belonging to the small chain are not counted as entry but kept in the sample as "Red stores" after the acquisition.

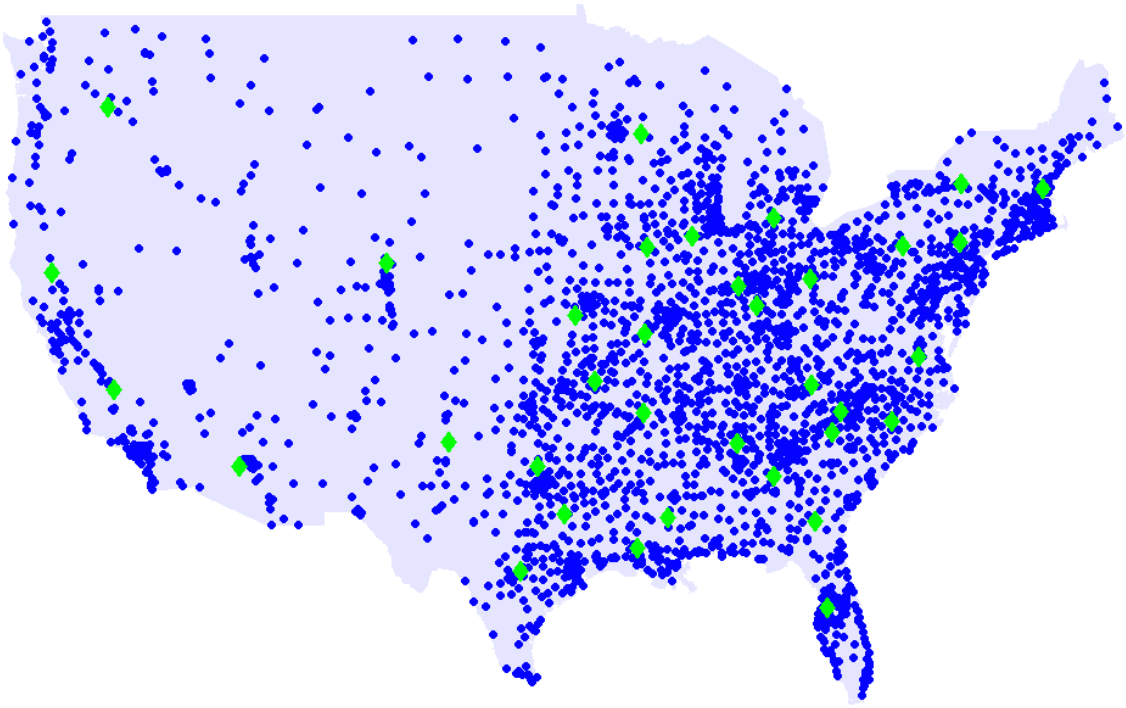

Figure 1.1: *Blue stores and distribution centers in 2001*

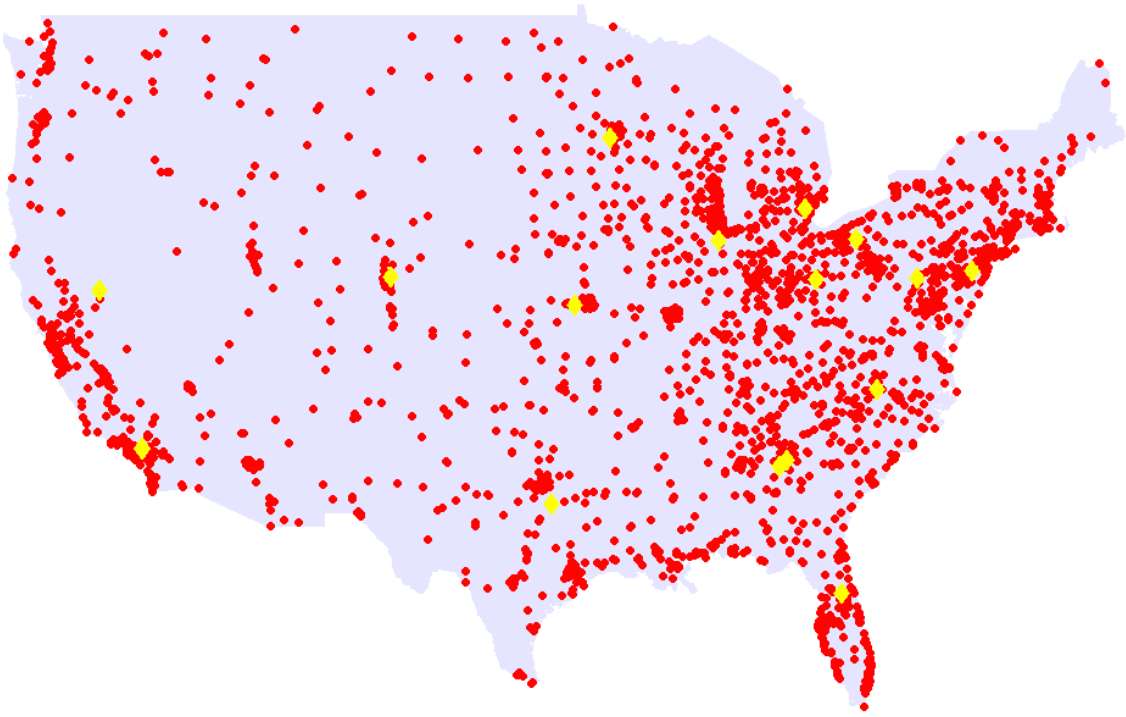

Figure 1.2: *Red stores and distribution centers in 2001*

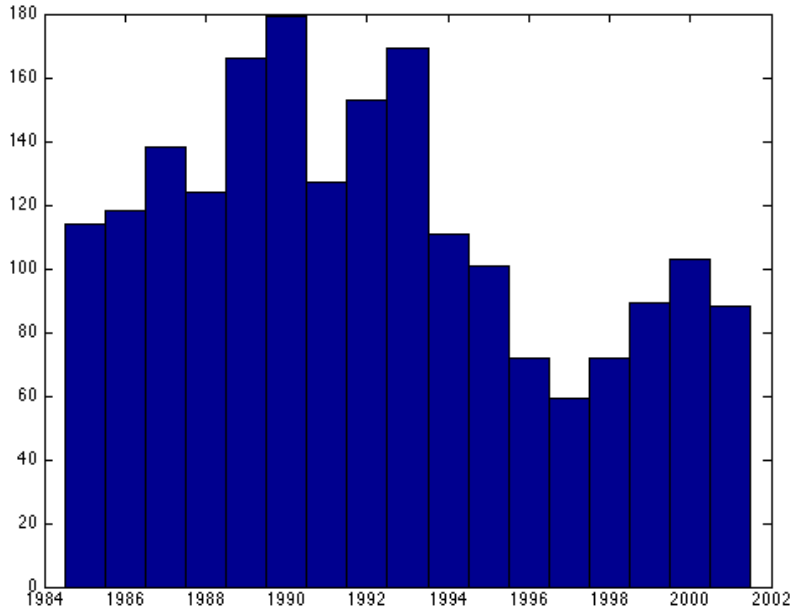

Figure 1.3: *Blue store openings by year 1985-2001*

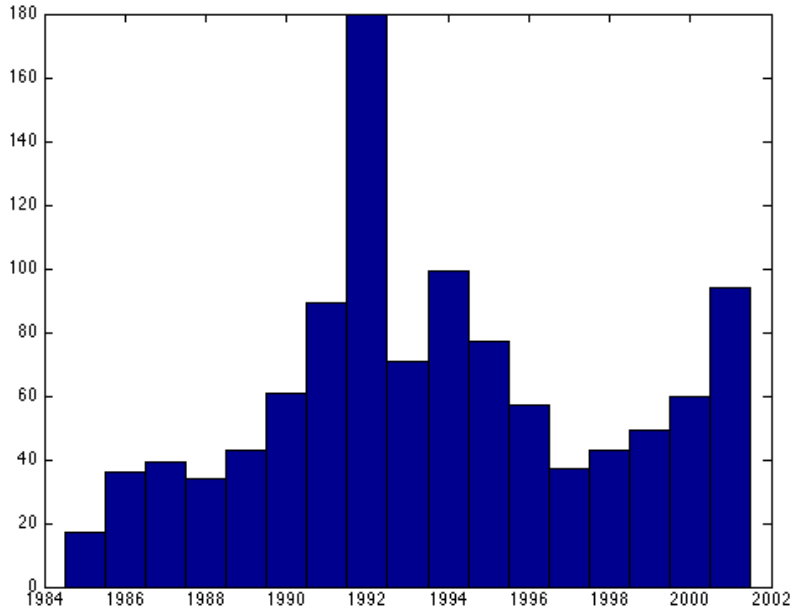

Figure 1.4: *Red store openings by year 1985-2001*

Table 1.2: *Location comparisons between two firms in sample period 1985-2001: median store characteristics measured in 2001*

	Blue	Red
distance to closest competitor's store	8.38	3.46
std. dev.	20.11	11.22
distance to closest same firm's store	11.90	10.16
std. dev.	15.12	20.79
total number of stores	1983	1140
number of same firm's stores in 30mi	4	3
number of any stores in 30mi	7	9
population density (10^5)	1.04	8.16
std. dev.(10^5)	3.20	1.86
distance to distribution center	98.47	126.56
std. dev.	71.41	122.02
total number of distribution centers	35	18

difference between Blue's median distance to the closest Blue firm, 11.90 miles, and that of Red, 10.16 miles, it appears that Blue stores are more spread out than Red stores. The number of any stores within 30 miles and population density around stores also indicate that Red stores are located in more concentrated areas, in terms of both store density and population density. Finally, Blue firm has 35 distribution centers while Red has 18. With more distribution centers, Blue stores are on average about 40 miles closer to their own distribution centers than Red stores to theirs. These differences, as will be discussed later, are important in characterizing preemptive entry behavior.

The second component of the data is store level characteristics. The store level sales

estimates and square footage of selling space of Blue and Red firm in 2007 come from the Nielsen TDLinx data. The sales are estimated using multiple sources including self-reported retailer input, store visits, questionnaires to store managers, etc.. They are regarded as the best available store level sales data of the discount retail industry and as a consequence they have been used by other researchers (Ellickson *et al.*, 2013, Holmes, 2011). For stores that sell both general merchandise and grocery, only sales of general merchandise are included. Square footage of selling space is derived from actual property plans. Because of the proprietary nature of this data set, I cannot present summary statistics of store characteristics.

The third component of the data consists of demographic information, wage, rent, and other information about the two firms. I use block group level demographic data in the 1980, 1990, and 2000 decennial censuses. A block group is a geographic unit that has a population between 600 to 3000 people. The demographic information of each block group contains total population, per capita income, share of African-American population, share of elderly population (65 years old and above), and share of young population (21 and below). Table 1.3 presents the summary statistics of the block group level demographic information. Wage data is constructed using average retail wage by county in the County Business Patterns between 1985 and 2003. Rent data is created using the residential property value information in the 1980, 1990, and 2000 decennial censuses. I adopt the same method as in Holmes (2011) to construct an index of property values. (See Appendix A of Holmes (2011) for details.) Firms' annual reports and interviews I conducted with managers and consultants also provide supporting information. Goodjobfirst.org is a website that collects government subsidy data published from various sources. The list of subsidies are incomplete, but it gives an idea about the scale of the subsidies. It is the best data source of its kind. Shoag and Veuger (2014) use this data in their study. I also interviewed a manager of Blue firm and a former manager of Red firm. I use the information collected from those interviews to choose between different modeling options, so that the model mimics how managers make decisions in reality.

Table 1.3: *Summary statistics of block group demographics*

	1980	1990	2000
mean population	0.83	1.11	1.35
mean income per capita	14.73	18.56	21.27
mean share African-American	0.10	0.13	0.13
mean share elderly	0.12	0.14	0.13
mean share young	0.35	0.31	0.31
no. of observations	269,738	222,764	206,960

source: U.S. census 1980, 1990, 2000

1.2.3 Descriptive Evidence of Preemptive Entry

In this section, I provide suggestive evidence of preemptive entry using reduced form regressions. Preemption, in this context, refers to the entry by one firm in order to deter entry by its opponent. More specifically, I define it as how much, in equilibrium, the likelihood of one firm entering a particular location today is impacted by the likelihood of its opponent entering the same location in the future, holding static profits constant. (A formal definition will be given in Section 2.3.) Using descriptive data, three kinds of behavior could be called preemption. First, a firm can open more stores than otherwise optimal, so that the opponent cannot enter the market. Second, a firm can cluster its stores to deter entry of the competitor. Finally, a firm can open a store earlier than otherwise optimal, so that the competitor cannot enter. The first two types are hard to find evidence for using reduced form regressions, since it is hard to separate store quantity and store density from unobserved market profitability. Therefore, I focus on the timing of store opening. More precisely, I choose to study Blue firm's store opening time instead of Red firm's, for two reasons. First, Blue firm's fast growth and high store density suggests it is more likely that it engaged in preemption, as described in the previous section. Second, during the observed

period of time, Blue firm has more observations of new stores openings than Red firm.

Next, I describe how preemptive incentives can be identified. The goal is to find a location characteristic that 1) affects Red firm's payoff of entering the location and therefore Blue firm's dynamic payoff if Blue firm does not enter the location in the current period, 2) does not affect Blue store's static profits of entry in the current period, i.e. is not correlated with unobserved market profitability. In other words, the impact of this location characteristic on Blue store's opening time should indicate how much the likelihood of Red store's entry affects Blue firm's entry decision.

One variable satisfying these conditions is the distance between a Blue firm's store and the closest Red firm's distribution center. The distance to the Red distribution center affects Red firm's payoff of locating a store, thus the likelihood of Red firm's entry. On the other hand, this distance does not directly impact Blue firm's static profit of entry. The challenge is that distribution centers are likely to be located close to potential stores, so that locations of Red distribution centers can be correlated with unobserved market profitability in the area. I include in the regression a control variable that approximates the unobserved market profitability around each Red distribution center. The profitability is measured by the total number of stores around the Red distribution center, including both Blue and Red stores, by the end of the observed period[10].

The Cox hazard model is applied to examine the impact of the distance to Red distribution center on Blue store's opening time. The dependent variable is duration before store opening for each Blue store l, measured in quarter. The observed time period is between 1985 and 2001. The independent variable of interest is the distance between l and the closest Red distribution center. Since Red firm was expanding its distribution center network during the period of observation, the distance to Red distribution center is time dependent. Thus each observation is a location l observed in period t. Let h_{lt} be the store

[10]The underlying assumption in the analysis is that the unobserved market profitability that is correlated with the locations of distribution centers does not fluctuate very much over time. This is likely to be true since all distribution centers are located in very rural and remote areas where demographics did not change very much over the sample period.

Table 1.4: *Evidence of preemptive entry: Control variables*

Blue firm's own store network and store density:
distance to the closest distribution center
distance to the closest Blue store
number of Blue stores within 30 and 50 miles
Competitor's store network and store density:
distance to the closest Red firm's distribution center
distance to the closest Red firm's store
number of Red stores within 30 and 50 miles
Location characteristics:
local wage and rent at time of opening
local population and demographics within 30 miles of the location

opening hazard rate of location l in period t.

$$ln(h_{lt}) = ln(h_{0t}) + \beta_1 d_l(t) + \beta_2 x_l'(t),$$

where $h_0(t)$ is the baseline hazard rate at time t, $d_l(t)$ is the distance between location l and its closest Red distribution center, and $x_l(t)$ is a set of control variables. The control variables include Blue firm's store and distribution center network characteristics[11], Red firm's store characteristics and other location characteristics such as wage, rent, and demographics. (See table 1.4 for detailed descriptions.) Since the stores in the sample are those that did eventually get opened, the regression captures the incentives for firms to manipulate the order of actions for strategic reasons.

[11]Since Blue firm was also expanding its distribution center network, distance to Blue's distribution centers is also time dependent.

Column 1 of Table 1.5 presents the results of the Cox hazard regression. Standard errors are clustered at the location level. The estimate indicates that a 100 mile increase in the distance between a Blue store and the closest Red distribution center reduces the hazard rate of Blue firm's store opening by 1.6%. Table 1.5 column 2 reports the same regression using an OLS framework. In this case, each observation is a store location. The estimated coefficient on distance to Red distribution shows that when the distance between a Blue store and its closest Red distribution center decreases by 100 miles, the opening time of the store becomes 1.2 quarters earlier on average. These results suggest that, if Red firm is also more likely to enter the same location, Blue firm is more likely to enter the location earlier than otherwise. This is suggestive evidence of preemption.

Table 1.5: *Evidence of preemptive entry: Blue firm's timing of store openings*

	Duration before store opening, 1985-2001	
	Cox Hazard Model	OLS
distance to closest red distribution center	-0.016	1.179
	(0.008)	(0.246)
tot. no. stores around red distribution center	0.002	-0.021
	(0.001)	(0.010)
distance to closest blue distribution center	-0.011	-0.470
	(0.002)	(0.435)
distance to closest blue store	0.533	-2.424
	(0.133)	(1.132)
no. of blue stores within 30mi	-0.048	0.694
	(0.016)	(0.171)
no. of blue stores within 50mi	-0.129	1.450
	(0.010)	(0.100)
distance to closest red store	0.632	-6.224
	(0.120)	(1.691)

Continued on next page

Table 1.5: *(continued)*

	Cox Hazard Model	OLS
no. of red stores within 30mi	0.041	-0.360
	(0.016)	(0.154)
no. of red stores within 50mi	-0.068	0.682
	(0.012)	(0.119)
local rent	0.000	0.002
	(0.000)	(0.003)
local wage	-0.078	0.644
	(0.016)	(0.150)
no. of blue stores by 2002	0.113	-1.278
	(0.008)	(0.078)
no. of red stores by 2002	0.023	-0.162
	(0.010)	(0.095)
N	61544	1983
R^2		0.66

1.3 Model

1.3.1 Overview

The model consists of two parts, a demand model and an entry model. The demand model is needed for computing sales of each existing and potential store. It includes detailed geographic information of consumer and store locations which allows store sales to be spatially interdependent. Another source of spatial interdependence across stores comes from the entry model. Entry decisions are modeled at the firm level subject to a budget constraint in a dynamic discrete-choice game framework. Then two-stage budgeting and separability are applied to make the model tractable. Finally, a clustering algorithm based

on separability conditions is employed to define markets. Section 1.3.2 describes the demand model, and Section 1.3.3 explains the entry model.

1.3.2 Demand

A demand model is needed in order to compute sales for each store location given consumer demographic information and location characteristics. There are two main ways to model demand in the literature. First, one can follow a Berry *et al.* (1995) type of model in which consumers in each market choose from the same set of products. Markets are independent and heterogeneity in consumer characteristics translates to different market shares of the same product in each market. This model allows for unobserved preference heterogeneity via random coefficients. Alternatively, one can adopt the demand model as in Holmes (2011). In this model, there is no market division and each consumer has its own choice set. This model has detailed geographic information about consumers and stores, and generates spatial interdependence across store locations. The drawback of this model is that it does not allow for unobserved preference heterogeneity due to the burden of computing a different set of choice probabilities for each consumer.

I choose the latter approach because spatial interdependence is important for modeling chain store's entry decision. Since payoffs are maximized at the firm level, when evaluating the payoff of a new store, firms need to take into account the impact of existing stores, including each firm's own stores and its competitor's stores. For example, if Blue firm is considering opening a new store in Boston, MA, it needs to evaluate the profitability of this location bearing in mind the existing stores in the neighboring town of Cambridge, since people living in both Boston and Cambridge can easily shop from all the stores located in either of the two cities.

The drawback of this approach is that it does not allow for unobserved preference heterogeneity. For example, the model is not able to capture the fact that different consumers dislike distance between home and stores with different intensities. In theory, random coefficients can be added to the model to account for unobserved heterogeneity. In practice,

however, it is computationally infeasible. The model is already difficult to estimate given that each unit of consumers (i.e. a block group) has a different choice set and a different set of choice probabilities and that there are over 200,000 block groups in the continental U.S. However, interaction terms in the regression and the definition of choice set for each block group can be used to mediate the problem. A detailed explanation is provided later in this section.

Each consumer i is a block group. Let u_{ijl} be the utility of consumer i shopping at firm j's store l.

$$u_{ijl} = \beta x_{jl} + \gamma_1 d_{il} + \gamma_2 d_{il} \times popden_i + \varepsilon_{ijl},$$

where x_{jl} is a vector of store characteristics including a brand dummy indicating if the store belongs to Blue or Red firm, the size of the store and if the store is newly opened in the current period. d_{il} is the distance between consumer i and store l. $popden_i$ is population density at block group i. Population density is measured by log of thousand people[12] within 5 miles of block group i. When population density varies, the interaction of distance and population density captures the heterogeneity in consumers' preferences with respect to distance to shops. ε_{ijl}'s are independent identically distributed and follow a type I extreme value distribution. Let u_i0 be consumer i's utility of shopping from an outside option, i.e. a store that does not belong to either Blue or Red firm:

$$u_{i0} = \alpha w_i + \varepsilon_{i0},$$

where w_i is a vector of covariates that include a constant, population density, population density squared, per capita income, and share of african-american, elderly, and young in the population. u_{i0} allows the utility of shopping from the outside option to depend on location characteristics. For example, more populated areas have more outside options and thus higher utility of not shopping from any Blue or Red stores in the choice set. This attempts to control for other competitors Blue and Red firms face in the market.

Block group i's choice set is defined as Blue and Red stores within r_i miles of the block

[12]Block groups with less than 1000 people are grouped together.

group. r_i is a function of population density,

$$25 \times (1 + (median(popden) - popden_i)/median(popden)).$$

r_i is in the interval between 17 and 35 miles, and equals 25 miles for the median block group with respect to population density. Letting r_i depend on population density captures the heterogeneity of consumer preferences towards distance between home and shop across areas with different population density, in terms of the furthest store they are willing to travel to. r_i increases as population density decreases. In other words, people living in rural area might be willing to travel further to a shop than those living in urban areas.

Let p_{ijl} be the probability of consumer i shopping at store l. Then store l's revenue is

$$R_{jl} = \sum_{i:d_{il} \leq r_i} \lambda \cdot p_{ijl} \cdot n_i, \tag{1.3.1}$$

where λ is average spending per consumer and n_i is the total population in block group i. Ideally λ might depend on consumer characteristics w_i. But the data is not detailed enough to identify $\lambda(w_i)$. This is because sales are only observed at the store level and that each store has a different set of consumers i patronizing it. One would need individual consumer level spending data to identify $\lambda(w_i)$. The constant λ is the average spending per consumer across the nation. In one of the empirical specifications, λ is allowed to depend on if store j sells general merchandize only or both general merchandize and grocery[13]. Results do not change very much. (See Section 2.2.1 for details.)

1.3.3 Firm's entry decision

In this section, I describe the firm entry model and show how it becomes tractable by applying two-stage budgeting and a clustering algorithm. I give an overview of the model in 1.3.3.1, present the details of the model in 1.3.3.2, discuss two-stage budgeting in 1.3.3.3, derive the separability conditions in 1.3.3.4, and explain the clustering algorithm in 1.3.3.5.

[13] As described in Section 1.2.2, the sales data of stores that sell both general merchandize and grocery only includes sales of general merchandize.

1.3.3.1 Overview of multi-store chain's entry model

There are three features of this multi-store chain's entry model. The first feature is that firms maximize payoffs over all stores instead of at each store independently. This is important because of the nature of multi-store retail chains, as well as of the spatial interdependence between stores as illustrated by the demand estimation in 2.2.1. The second feature is that firms are forward-looking. This is necessary when examining preemptive incentives. Given that demographics and distribution networks are changing over time, it is reasonable to assume that firms maximize the sum of expected current and future payoffs. Holmes (2011) also showed that dynamic consideration is important for discount retail chain's entry decisions. The third feature is that there are strategic interactions between firms. This is also necessary for studying preemptive entry. It is supported by the fact that Blue and Red firm compete in many markets as shown in Figure 1.1 and Figure 1.2. Findings in Jia (2008) provide more evidence of strategic interactions. Therefore, I model firms' entry decisions using a discrete choice game framework in which decisions are made at the firm level.

In each period, firms choose the locations of a set of new stores to maximize the current profits and the sum of discounted future values. When making the decision, each firm takes into account the current and future demographics, distribution networks, local wage and rent, its own store openings in the future, and its opponent's store openings in current and future periods. The decision is made at the firm level instead of individual store level. The number of new stores to be opened is determined by a budget constraint. Since both firms were expanding in the sample period, they face financial constraints. Moreover, the constraint is necessary for studying preemptive incentives. Firms move sequentially each period. Blue firm moves first.

The large number of possible locations and store openings of both firms in each period leads to the very large state space in the game. As a result, the firm optimization problem has high computational complexity for the firms as well as for the econometrician. Tools that make the model tractable mimicking the way firms solve the problem in reality are therefore desirable.

1.3.3.2 Two-player discrete choice game

Let π_{jt}^l be the static profit of firm j's store l in period t.

$$\pi_{jt}^l = \mu_j R_{jt}^l(s_t) - w_t^l E(R_{jt}^l(s_t)) - r_t^l L(R_{jt}^l(s_t)) - \psi_j D_{jt}^l - \alpha_j x_{jt}^l, \tag{1.3.2}$$

where μ_j is the gross margin of firm j. $s_t = (s_{jt}, s_{-jt})$, and $s_{jt} = \{0,1\}^L$ indicating if j has a store at each location of all possible locations $\{1, ..., L\}$. Each location l contains up to one store. Denote j's opponent by $-j$. $R_{jt}^l(s_t)$ is the revenue of store l in period t which depends on s_t, the locations of both j's stores and j's opponent's stores. I follow Holmes (2011) in modeling labor cost and land cost as variable cost. w_t^l and r_t^l are local wage and rent. $E(R_{jt}^l(s_t))$ is the number of employees and $L(R_{jt}^l(s_t))$ is the size of land. $\psi_j D_{jt}^l$ is the distribution cost where ψ_j is per unit distribution cost and D_{jt}^l is distance to distribution center. $\alpha_j x_{jt}^l$ is fixed cost which depends on population density around store l, x_{jt}^l. Each period t is a quarter. The firm level static profit is

$$\pi_{jt} = \sum_{l=1}^{L} s_{jt}^l \left\{ \mu_j R_{jt}^l(s_t) - w_t^l E(R_{jt}^l(s_t)) - r_t^l L(R_{jt}^l(s_t)) \ - \psi_j D_{jt}^l - \alpha_j x_{jt}^l \right\}, \tag{1.3.3}$$

where the sum is over all the locations of firm j stores.

Next I introduce the value function of the firm. For simplicity, I assume sequential move and that Blue firm moves first[14]. a_{jt}^l denotes firm j's action at location l in period t, where $l \in \{1 \cdots L_t\}$. $a_{jt}^l = 1$ if j opens a new store at l in period t, and $a_{jt}^l = 0$ otherwise. L_t is the set of all possible locations minus those taken by the two firms before period t, i.e. $L_t = L/\{s_{jt}, s_{-jt}\}$. $s_{jt+1} = s_{jt} + a_{jt}$, where $s_{jt} \in \{0,1\}^L$. Let z_{jt} be the location of j's distribution centers in period t and B_{jt} be the budget constraint of firm j in period t. For notational simplicity, let $s_{jt} = (s_{jt}, z_{jt}, B_{jt})$, and $s_t = (s_{jt}, s_{-jt})$ be the state variable. Firm j's

[14]This is a strong assumption. In the future, I would like to flip the order and solve the game for when Red firm moves first.

value function in period t is

$$V(s_{jt}, s_{-jt}) = \max_{a_{jt} \in \mathbb{A}_t} \left\{ \mathbb{E}\pi(s_{jt} + a_{jt}, s_{-jt}) + \beta \sum_{s_{-jt+1}} \mathbb{E}V(s_{jt} + a_{jt}, s_{-jt+1}) P(s_{-jt+1} | s_{jt+1}, s_{-jt}) \right\} \tag{1.3.4}$$

s.t.

$$\sum_{l=1}^{L_t} f(a_{jt}^l) \leq B_{jt}, \tag{1.3.5}$$

where $\mathbb{A}_t = \{0,1\}^{L_t}$ is the choice set in period t subject to the budget constraint B_{jt}. The expectation is over a cost shock η_{jt}^l of entering at location l. η_{jt}^l are i.i.d. across locations and time periods. $P(s_{-jt+1}|s_{jt+1}, s_{-jt})$ is the transition probability of j's opponent in period t. $f(a_{jt}^l)$ is the budget function which I will discuss in more detail. β is the discount factor.

Each period, firm j chooses the optimal entry decision $a_{jt} \in \mathbb{A}_t$ to maximize the sum of expected profits $\mathbb{E}\pi(s_{jt} + a_{jt}, s_{-jt})$ and the continuation value $\beta \sum_{s_{-jt+1}} \mathbb{E}V(s_{jt} + a_{jt}, s_{-jt+1}) P(s_{-jt+1}|s_{jt+1}, s_{-jt})$. The distribution of η_{jt}^l is common knowledge but its realization is private information. Firm j's strategy σ_j is a function from the state variable s_t to a set of choice probabilities $Pr(a_{jt}|s_t)$. Perception of future states $P(s_{t+1}|s_t)$ is consistent with equilibrium play. The solution is a Bayesian Markov Perfect Equilibrium.

The difference between this model and the commonly used incomplete information dynamic game framework (Ryan, 2012) is that it has a budget constraint in Equation (1.3.5). Think of the budget function f as a cost of opening $\sum_l a_{jt}^l$ stores in period t. It represents the actual costs of acquiring land or building infrastructure, as well as the management costs of hiring workers or submitting paperwork to the local government. There are three reasons to include this condition. First, it mimics the way firms behave. According to the managers I interviewed, each period, firms designate a certain amount of funds for the opening of new stores, which is equivalent to a budget constraint. Second, both firms are expanding in the sample period 1985-2001. Financial constraints can often be a serious consideration when firms are expanding. Figure 1.5 plots the book value of total assets of Blue and Red firm in the sample period, in respective colors. The figure shows that Blue firm's book value of total assets grew fast in this period. Thus it is likely that the financial constraints

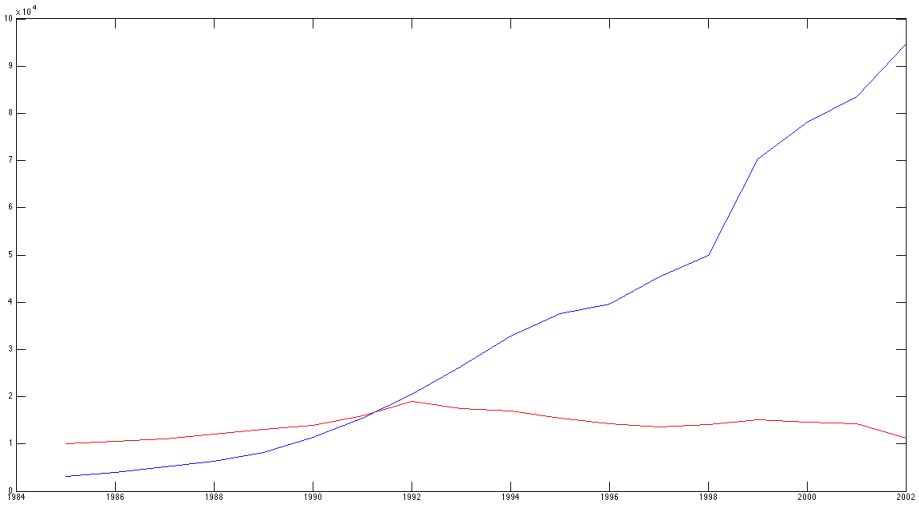

Figure 1.5: *Book value of total assets, 1985-2002*

Blue firm is facing are substantial at the beginning of this period and are reduced by the end of this period. Third, the budget constraints are necessary for studying preemptive incentives. If firms were not liquidity constrained, in theory, they could enter all markets to deter entry by the competitor in period 0. This is unrealistic both in terms of financial costs and management costs. For simplicity, I assume hereafter that $\sum_{l=1}^{L_t} f(a_{jt}^l) = \sum_{l=1}^{L_t} a_{jt}^l \leq B_{jt}$, i.e. the total number new stores each firm can open in each period is held fixed at the observed level. Note that although B_{jt} is a choice made by the firm, the assumption does not cause selection issues since the choice probabilities become conditional probabilities given the optimal budget constraint B_{jt}.

Next I explain how the set of potential locations L is defined in the game. I restrict L to be all the locations Blue and Red firm eventually entered by the end of the sample period. The alternative would be to include all possible locations regardless of the existence of a store at any point in time. There are two reasons for choosing the former approach over the alternative one. First, for the purpose of studying preemptive incentives, it is reasonable to focus on locations firms are potentially interested in entering. If a location is very far from being profitable enough for either firm to ever enter, it does not provide information

for identifying preemptive incentives. Second, including locations where no entry is ever observed implies dividing the U.S. national market into many smaller markets. The division usually involves using census geographic units as markets (Jia, 2008, Zhu and Singh, 2009, Ellickson *et al.*, 2013). This allows little spatial competition and, as shown below, may lead to biased results. The drawback of defining observed stores by the end of the sample period as the set of potential locations is that firm's decisions are very much affected by the limited choice set towards the end of the sample period. To avoid the problem, I leave out the last two years of data and only include observations between 1985 and 1999 as the sample of study. The choice of leaving out two years, specifically, will be explained in 2.2.3.

Modeling entry decisions at firm level and allowing for spatial interdependence of store locations captures the nature of spatial competition between multi-store chains, but it also makes the problem intractable. Each period, firms are choosing from the set of potential locations that have not been occupied. Since both firms are expanding very fast in the sample period, the choice set is very large. Take $t = 36$, the fourth quarter of 1993, as an example, Blue firm and Red firm opened 27 and 24 stores respectively. The total number of potential locations is 1262. With a total of 1262 locations, the size of the state space is $\binom{1262}{27} \approx 10^{35}$.

1.3.3.3 Two-stage budgeting

In this section, I describe how two-stage budgeting and separability can be applied to make the model tractable, while retaining the features described above. Two-stage budgeting refers to the fact that consumers first allocate a given amount of total expenditure to categories of goods and then optimize consumption within each category, conditional on the amount of expenditure designated to this category of goods (Gorman, 1971)[15]. I apply this idea to chain stores' entry problem. Store locations are similar to goods in the consumption problem. Let $\{1, ..., P_{M_t}\}$ be a partition of potential locations $\{1, ..., L_t\}$ in period t. Partitions mimic the categories of goods in the consumption problem. Two-stage budgeting implies

[15]Note only the separability conditions are needed here, the conditions for constructing price index are not.

that firms solve the following problem. For each $P_m \in \{1, ..., P_{M_t}\}$,

$$V(s_{jmt}, s_{-jmt}|B_{mt}) = \max_{a_{jmt} \in \mathbb{A}_{mt}} \Bigg\{ \mathbb{E}\pi(s_{jmt} + a_{jmt}, s_{-jmt}) + \beta \sum_{s_{-jmt+1}} \mathbb{E}V(s_{jmt} + a_{jmt}, s_{-jmt+1}|B_{mt}) \cdot P(s_{-jmt+1}|s_{jmt+1}, s_{-jmt}) \Bigg\} \quad (1.3.6)$$

s.t.

$$\sum_{l \in m} a^l_{jt} \leq B_{jmt},$$

where $s_{jmt} = \{0,1\}^m$ and B_{jmt} is the budget constraint of element P_m of the partition $\{1, ..., P_{M_t}\}$. 1.3.6 corresponds to the consumption problem in stage 2. Then, firm j solves for the optimal budget $\{B_{j_1 t}, ..., B_{j_{M_t} t}\}$ for each element P_m of the partition:

$$\sum_{m=1}^{M_t} \mathbb{E}V(s_{jmt}, s_{-jmt}|B_{jmt}) \quad (1.3.7)$$

s.t.

$$\sum_{m=1}^{M_t} B_{jmt} \leq B_{jt},$$

which corresponds to the stage 1 of the consumption problem.

With two-stage budgeting, firms solve two smaller optimization problems, (1.3.6) and (1.3.7) instead of the problem in (1.3.4). This decentralization greatly reduces the size of the state space in estimation. When estimating a model like (1.3.6), one can condition on the the optimal budget of each element of the partition $\{B_{jmt}\}_{m=1}^{M_t}$, and only solve the stage 2 problem (1.3.6). The state space is then reduced to $\sum_{m=1}^{M_t} \binom{|P_m|}{\sum_{l \in m} a^l_{jt}}$.

Moreover, two-stage budgeting is a good approximation to how firms actually behave. Blue firm, for example, divides the U.S. national market into regions. According to the manager I interviewed, regional managers choose a set of potential new store locations each period, and submit the locations to the headquarter. Managers in the headquarter then rank the potential locations from all regions and decide which stores will be opened subject to a budget constraint.

However, it is not clear whether solving (1.3.6) and (1.3.7) is equivalent to solving (1.3.4).

In the next two sections, I show that under a set of conditions, namely separability, solving the two-stage budgeting problem in (1.3.6) and (1.3.7) is equivalent to solving the overall optimization problem (1.3.4). Then I derive sufficient conditions on the primitives of the model such that the separability conditions are satisfied. I define a market to be an element of the partition P_m. I show that the solution to the two-stage budgeting problem is optimal if separability across markets holds.

1.3.3.4 Separability conditions in a two-player Markov game

Separability is defined following the work of Gorman (1959, 1971) and is further generalized to be applicable to a two-player dynamic game setting. First, to build intuition, I define separability in the static game context. For simplicity, the subscript t is suppressed. Let σ_j be firm j's strategy. $s = (s_j, s_{-j})$ is the state variable. Firms solve

$$max_{\sigma_{j(s)}} \pi(s_j + a_j, s_{-j} + a_{-j}) \text{ s.t. } \sum_{l=1}^{L} a_j^l \leq B_j. \tag{1.3.8}$$

Let $\{P_1, \cdots, P_M\}$ be a partition of the potential store locations $\{1, \cdots, L\}$. Denote $\pi(s_j^l = 1, s_j^{-l}, s_{-j})$ the profit of j when j has a store at location l. Define

$$\Delta\mathbb{E}\pi(s_j, s_{-j}, l) = \mathbb{E}[\pi(s_j^l = 1, s_j^{-l}, s_{-j}) - \pi(s_j^l = 0, s_j^{-l}, s_{-j})],$$

to be the expected marginal profit of j entering l when the state variable s equals (s_j, s_{-j}), where the expectation is taken over the cost shock η_j^l.

Definition 1.1 *Locations $\{1, \cdots, L\}$are separable in the partition $\{P_1, \cdots, P_M\}$ if*

$$\frac{\Delta\mathbb{E}\pi(s_j, s_{-j}, l)}{\Delta\mathbb{E}\pi(s_j, s_{-j}, h)} \perp (s_j^k, s_{-j}^k), \forall l, h \in P_{m_l}, \forall k \notin P_{m_k},$$

where l, $h \in P_{m_l}$, and $k \in P_{m_k}$.

In other words, if the ratio of the expected marginal profits of opening any two locations in a market does not depend on the state variables in another market, locations are separable with respect to markets. This too is analogous to the consumption problem, in which

separability holds when the rates of substitution of any two goods are independent across categories of goods (Gorman, 1959). Next, I define separability in strategy σ_j. Note $\sigma_j(s)$ can be written as a vector $(\sigma_j^1(s), \cdots, \sigma_j^M(s))$ for any partition $\{P_1, \cdots, P_M\}$. Similarly, any state variable s_j can be written as a vector $(s_{j1}, \cdots, s_{jM})$. Let σ_j^* be the best response of j given opponent's strategy σ_{-j}, and a_j^* be the corresponding optimal action at state (s_j, s_{-j}).

Definition 1.2 *Firm j's strategy σ_j^* is separable in the partition $\{P_1, \cdots, P_M\}$ if for given σ_{-j}, $\exists \sigma_{j1}^*, \sigma_{j2}^*, \cdots \sigma_{jM}^*$ s.t.*

$$\sigma_{jm}^*(s_{jm}, s_{-jm}, B_m) = \sigma_j^{*m}(s_j, s_{-j}, B),$$

where $\sigma_j^* = \left(\sigma_j^{*1}, \cdots, \sigma_j^{*M}\right)$, $B_m = (B_{jm}^*, B_{-jm})$, $B_{jm}^* = \sum_{l \in P_m} a_j^{*l}$, $B_{-jm} = \sum_{l \in P_m} a_{-j}^l$, $\forall m = 1, \cdots, M$, and $\sum_{m=1}^M B_m = B$.

In other words, σ_j^* is separable if each of its component σ_j^{*m} can be written as a function σ_{jm}^* which only depends on the state variable in the partition m, (s_{jm}, s_{-jm}), and on the budget constraint of the partition, B_m. This implies that conditional on the optimal budget of the partition B_{jm}^*, j is able to compute the best response in partition j with information within the partition m only, regardless of the values of state variables or budget levels in other components of the partition.

Theorem 1.1 *If locations $\{1, \cdots, L\}$ are separable in partition $\{P_1, \cdots, P_M\}$, and the opponent's strategy σ_{-j}^* is separable, then j's optimal strategy σ_j^* is separable.*

See Appendix I for the details of the proof. Theorem 1.1 states that if locations $\{1, \cdots, L\}$ are separable and that one firm is playing a separable strategy, then it must be optimal for the other firm to play a separable strategy as well. In other words, both firm's strategies are separable in equilibrium. Define such an equilibrium as separable equilibrium. Separable equilibrium is a refinement of Nash equilibrium.

Next, I derive sufficient conditions on the primitives such that separability of locations holds. There are four parts of the profit function (1.3.3) that need to be examined for separability. The first three terms in the profit function all depend on revenue $R_j^l(s_t)$, thus

$R_j^l(s)$ needs to satisfy the separability condition. The other three terms are the distribution cost, the fixed cost, and the cost shock when opening a new store η_j^l.

Theorem 1.2 *The location $\{1, \cdots, L\}$ is separable in partition $\{P_1, \cdots, P_M\}$ if the profit function $\pi(\cdot)$ satisfies the following conditions,*

1. *$R_j^l(s)$ is additively separable in partition $\{P_1, \cdots, P_M\}$,*
2. *Distribution cost, as well as fixed cost, at location l is independent of z_j^k and x_j^k, where $k \in P_n$,$m \neq n$,*
3. *η_j^l are independently distributed across markets.*

See appendix for the proof. By Equation (1.3.1), it is clear that if $\nexists i$, s.t.

$$p_{ijl} > 0, p_{ijk} > 0, l \in P_m, k \in P_n, m \neq n, \tag{1.3.9}$$

then the first condition in Theorem 1.2 holds. In other words, if there does not exist consumer i that shops from both store l in market P_m and store k which belongs to a different market P_n (i.e. stores in different markets do not share customers), then stores $\{1, \cdots, L\}$ are separable in partition $\{P_1, \cdots, P_M\}$. The second condition is automatically satisfied by the specification of distribution cost $\psi_j D_j^l$ and fixed cost $\alpha_j x_j^l$. The condition can be violated if, for example, the distribution center has a capacity constraint and per unit cost of distributing depends on the number of stores the distribution center serves. The third condition is satisfied by the i.i.d. assumption on the cost shock η_j^l.

Finally, I generalize the separability conditions derived above to a dynamic game setting with Bayesian Markov perfect equilibrium. The definitions are very similar to the static case, except for two differences: 1) the expected static profit function $\mathbb{E}\pi(s)$ becomes the expected value function $\mathbb{E}V(s_{jt}, s_{-jt})$ in Equation (1.3.4), 2) instead of the market level budget constraint in one period, strategies are separable conditional on the sequence of market level budget constraint $\{B_{jmt}, B_{-jmt}\}_{j=1}^{\infty}$, for all m. The results in Theorem 1.1 and Theorem 1.2 apply. See Appendix I for details and the proofs.

1.3.3.5 Separability and market division

In order to apply two-stage budgeting, a sufficient condition is that the markets are separable, as discussed in the previous section. In this section, I present the tools needed to divide the U.S. national market into smaller separable markets. This is done by applying a clustering algorithm based on the separability condition and the demand model. I first define the objective function for the clustering algorithm and then I explain the steps of the algorithm to find an optimal partition of store locations given the objective function. Each element of the partition is defined as a market. Results are postponed to Section 2.2.2.

The main condition that needs to hold for markets to be separable is that revenue is independent across markets. In other words, stores in two different markets do not share customers. This condition is automatically satisfied if two stores are so far away from each other, that no consumer has both stores in its choice set. Clearly, the difficulties in dividing markets arise when two markets are next to each other and consumers living close to the border of the two markets are willing to shop from either store. In reality, two neighboring stores almost never share no customer, except in areas where population density is extremely low and the stores are very far from each other. Therefore, I define an objective function for the clustering algorithm that captures how far away the partition is from being truly separable.

Define the objective function as the following,

$$min_{\{P_1,\cdots,P_M\}} \sum_{l=1}^{L_t} [R^l(s,\omega) - R^l(s_m,\omega_m|l \in P_m)]^2, \tag{1.3.10}$$

where L_t[16] is the set of potential locations in period t, $R^l(s,\omega)$ is the revenue of store l, which depends on the set of existing stores of both firms s and the determinants ω of demand which include demographic characteristics and store characteristics. Note both s and ω are vectors that contain information about the entire U.S. market. The second term $R^l(s_m,\omega_m|l \in P_m)$ is also store l's revenue, but it is computed using information of existing

[16]I kept the t subscript to differentiate L_t from L, which is all locations including both L_t the potential locations and those that have been entered up to period t.

store locations and demand data only in the partition P_m. That is, it is the revenue of store l if l is assigned to market P_m. In this case, some of the spatial interdependence between l and any other store h that belongs to a different market P_n, $n \neq m$, is not accounted for. In other words, if l and h share any customers, those customers are restricted to shop at only one of the two stores. Consumers in block groups that have both l and h in their choice set are assigned to the market that their closest store belongs to[17]. If P_m is truly separable from the rest of the markets, then $R^l(s,\omega) - R^l(s_m, \omega_m | l \in P_m)$ is zero for all $l \in P_m$. Therefore, the sum of squared differences between $R^l(s,\omega)$ and $R^l(s_m, \omega_m | l \in P_m)$ indicates how far off a partition is from each of its elements being truly separable, or the loss of assuming the elements are separable. The solution to Equation (1.3.10) finds the optimal partition that minimizes this loss.

Next, I introduce the clustering algorithm that attempts to find a solution to Equation (1.3.10). Since it is a graph partitioning problem that is NP-hard (Fortunato and Castellano, 2009), the solution is an approximation. Although there might be other approximated solutions to this problem, results in section 2.2.2 indicate that the clustering algorithm does reasonably well.

Start with $M = 2$. Apply a greedy algorithm which locally minimizes the objective function to find an approximated global solution to Equation (1.3.10). Then increase M and repeat the previous step. Stop when the stopping criterion binds. Due to the complex geographic structure of the model, greedy algorithm is more suitable than other algorithms such as spectrum algorithm that has the advantage of speed but assumes additional structure of the problem. I describe the greedy algorithm and the stopping criterion in the remainder of this section.

The greedy algorithm finds the optimal partition $\{P_1, \cdots, P_M\}$ given the objective function (1.3.10) and the number of clusters M. Figure 1.6 demonstrates the idea for $M = 2$. Each dot is a store. The edge between a pair of dots means that the two stores share customers. The task is to cut off a few edges such that the set of locations is divided

[17] I also tried assigning consumers to the market their most preferred store belongs to. Results are similar.

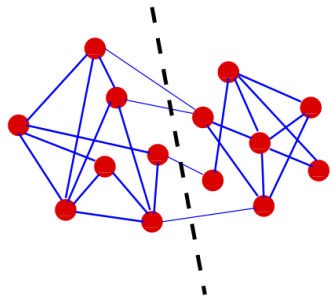

Figure 1.6: *Graph partitioning*[a]

[a] Fortunato and Castellano, 2009

into two markets. The broken edges are selected so that the objective function (1.3.10) is minimized. There are two features of the problem that are important for the set-up of the greedy algorithm. First, only stores close to the border of two markets matter. The objective function is zero for stores that have all connected neighbors in the same market. This feature leads to the fact that the algorithm focuses on stores close to the borders of markets. Second, the edges between stores are weighted. The weight is the amount of interdependence between two stores and is decided by the demand data and the objective function. The weight varies across edges, thus one cannot simply minimize the number of broken edges in a graph to find the optimal partition.

The algorithm follows four steps.

1. For a given partition of locations $\{P_1, \cdots, P_M\}^t$, find all l s.t. $\exists h \in C_l$, and $h \in P_n$, but $l \in P_m$, and $m \neq n$, where C_l is the set of locations l is connected to.

2. Reassign each l in the previous step to a partition such that (1.3.10) is minimized keeping the assignment of all the other stores fixed. Call the new partition $\{P_1, \cdots, P_M\}^{(t+1)}$.

3. Repeat step 1 and 2 untill the algorithm converges.

4. Repeat step 1, 2, and 3 for 1000 different initial partitions $\{P_1, \cdots, P_M\}^0$.

One nice property of this algorithm is that the resulting partition respects contiguity. If all the connected neighbors of some store(s) belong to one partition, the store(s) itself cannot

be in a different partition. I.e. If $h \in P_m$, $\forall h \in C_l$, then it must be that $l \in P_m$.

Finally, I explain the stopping criterion for picking the number of partitions M. As the number of clusters increases, the incremental change of loss, i.e. the value of (1.3.10), also increases. The stopping criterion is chosen when the sum of incremental change of loss from M to $M+1$ partitions is bigger than or equal to 1% of revenue $R^l(s,\omega)$ for any store l[18]. It also happens to be the point at which the incremental change of loss increases dramatically in many cases.

[18]Sensitivity checks on the 1% criterion are to be conducted. Future research is needed to obtain a more systematic stopping criterion.

Chapter 2

Spatial Competition and Preemptive Entry in the Discount Retail Industry: Structural Estimation

2.1 Introduction

Following the behavioral model described in Chapter 1, this chapter provides empirical evidence to preemptive entry and evaluates local government's subsidy policy in the discount retail industry.

First, this chapter demonstrates how the cost parameters are estimated using the behavioral model and the data described in Section 1.2.2. This is done in three steps. In the first step, consumer preferences are recovered by a demand estimation that allows consumers to shop from nearby stores and therefore stores to compete for the shopping dollars of the consumers. In the second step, potential store locations that firms are choosing from are divided into markets. Market divisions are inferred from data and the estimated consumer preferences. In the last step, cost parameters are recovered by solving a two-player dynamic location game. The continuation values of the potential store locations are computed using a rolling-window approximation.

Using the cost estimates, I conduct two counterfactual studies. First, to quantify preemptive incentives in firms' store location decisions, I compute the optimal decision firms would have made if the amount of preemptive incentives they are facing is reduced in the game. By computing the sum of discounted current and future payoffs of both firms, I show that preemption leads to loss of producer surplus of about $1 million dollars per store on average. Second, I compute the amount of subsidies Blue firm would need to open stores at those locations Red firm exited and show that the observed subsidy level is too low to affect Blue firm's entry decisions.

The chapter is organized as follows. Section 2.2 shows how the value functions can be approximated and presents the estimation results. Counterfactuals under which preemptive motives are removed are presented in Section 2.3. The subsidy policy application is presented in Section 2.4. Section 2.5 concludes.

2.2 Estimation and Clustering

2.2.1 Demand estimation

I estimate the demand model using a maximum likelihood framework. Like Holmes (2011), I assume the discrepancy between the model and the data is measurement error and that the error follows a normal distribution. Denote the measurement error by ϵ_{jl} and observed sales of store l by R_{jl}^{obs}. Then

$$ln(R_{jl}^{obs}) = ln(R_{jl}) + \epsilon_{jl},$$

where $\epsilon_{jl} \sim N(0, \sigma^2)$.

Results of the demand estimation are presented in Table 2.1. The first column shows the results of the basic specification. Spending per person is on average $47 per week, which is $2444 per year. This is close to the estimate in Holmes (2011), that is of about $2150 per year in 2007 dollars. The coefficient of population density is positive and significant. The coefficient on distance is negative and significant. I interpret the coefficients using comparative statics in Table 2.2. Column 2 presents a different specification with a dummy

variable indicating whether a store sells both general merchandise and grocery. The results are similar.

Table 2.1: *Demand estimates*

Average weekly sales in \$1000*s* by store

	specification1	specification2
σ^2	0.082	0.080
	(0.002)	(0.002)
spending per person	0.047	0.050
	(0.002)	(0.002)
grocery dummy		-0.005
		(0.000)
constant	-3.149	-3.151
	(0.253)	(0.269)
popden	1.270	1.270
	(0.060)	(0.062)
$popden^2$	-0.022	-0.020
	(0.006)	(0.006)
per capita income	0.009	0.010
	(0.002)	(0.002)
black	0.246	0.318
	(0.062)	(0.063)
old	-0.522	-0.566
	(0.279)	0.285
young	0.052	0.052
	(0.336)	(0.344)
size	0.504	0.505
	(0.198)	0.200

Continued on next page

Table 2.1: *(continued)*

	specification1	specification2
$blue$	0.312	0.312
	(0.053)	(0.053)
new	-0.117	-0.107
	(0.024)	(0.022)
$distance$	-0.440	-0.441
	(0.020)	(0.021)
$distance * popden$	0.020	0.019
	(0.005)	(0.005)
R^2	0.843	0.845

Number of stores=4750

Number of blockgroups=202020

Standard errors are in parenthesis.

I use comparative statics to illustrate the effects of distance to shops in the choice set and population density on store sales. This exercise also demonstrates how spatial competition is generated from demand. Consider a Red store located two miles away from the median block group and a new Blue store entering the market. First, I fix the population density of the block group and compute the probabilities of consumers shopping at the Red store when the distance between the new Blue store and the block group changes. Column 1 in Table 2.2 reports the choice probabilities when population density equals 1. Moving up across the rows, the choice probabilities decrease as distance to the Blue store decreases. This shows that the competition between the two stores intensifies as the Blue store moves closer to the consumers. The result stays the same across columns when population density takes on different values. Row 1 reports the probabilities of consumers shopping at the Red store when population density increases and distance to Blue store stays at 2 miles. From left to right, choice probabilities decrease as population density increases. This shows that

Table 2.2: *Demand comparative statics*

Probabilities of shopping at Red store				
	Population density			
Distance	1	10	50	250
2	0.18	0.14	0.06	0.01
4	0.34	0.21	0.07	0.02
10	0.80	0.33	0.08	0.02
20	0.88	0.34	0.08	0.02

the utility of choosing the outside option increases as population density increases.

2.2.2 Clustering results

In this section, I present the clustering result and compare it to an alternative definition of markets in the literature, Core Business Statistical Areas (CBSA).

Since the set of potential locations L_t changes every period, the partitioning of markets is conducted every period. Note the existing stores are not partitioned because they are no longer in the choice set of the firms. Consequently, the spatial interdependence between existing stores and potential stores is fully accounted for. As a result, in the empirical analysis, market definitions are different in each period. The advantage of this assumption is that it is close to how managers actually make entry decisions. According to the managers and consultants I interviewed, when firms evaluate a potential store location, they define a trade area around it. The trade area is defined as the area which demand is likely to come from and which the main competing stores including the firm's own stores and competitors' stores are located at. Naturally, the trade area varies across time as new stores are opened each period. Therefore, the clustering procedure can also be viewed as estimating trade areas each period. The disadvantage of doing so is that it does not allow any market level

unobservables that can be controlled for using market fixed effects.

Figure 2.1 is a map of the northeast, mid-atlantic and (part of) southeast United States. It is the area where store density was the highest in the third quarter of 1997. The points (squares, dots, and diamonds) are the potential store locations in this period. Neighboring stores are marked with different colors or shapes to display market divisions. The CBSAs also appear in the map, in yellow. This map compares market divisions defined by the clustering algorithm to the CBSA units. In a few cases, using CBSA as a market is not very different from the clustering of the store locations. For example, the single black dot in the very northeast is the only store in its market after clustering. In this particular case, defining market as the CBSA of the store may not be a bad idea, since there are no other store around. However, in most cases, defining a CBSA as a market can be misleading. For example, the five blue dots just southwest of the black dot are defined as in one market by clustering, while they are located in three different CBSAs. Dividing the stores into three different market would be misleading since they are very close to each other, and at least two of them are almost right at the border of two CBSAs. Consumers do not confine themselves to shop within CBSA boundaries, so it is not reasonable to define markets with this restriction. Such restriction is minimized in the division of markets by applying the clustering algorithm.

Table 2.3 presents a few measures demonstrating the goodness of the clustering results and compares them with those by using CBSAs as markets. The measures are computed using all L_t locations in the third quarter of 1997. There are 241 potential locations. First, the total loss by clustering as a fraction of total sales, i.e. $[\sum_l |S(l) - S(l \in r^*)|]/[\sum_l |S(l)]$, is less than 0.001%. The maximum store level loss as a fraction of sales is also reasonably small, 0.5%. Finally, the total number of stores affected by clustering is 55. This shows that simply excluding these locations is not a satisfying option, since it reduces the sample size by more than 20%. On the other hand, using each CBSA as a market would lead to undesirable results. The maximum store level loss is 53.0%, about a hundred times higher than that of clustering.

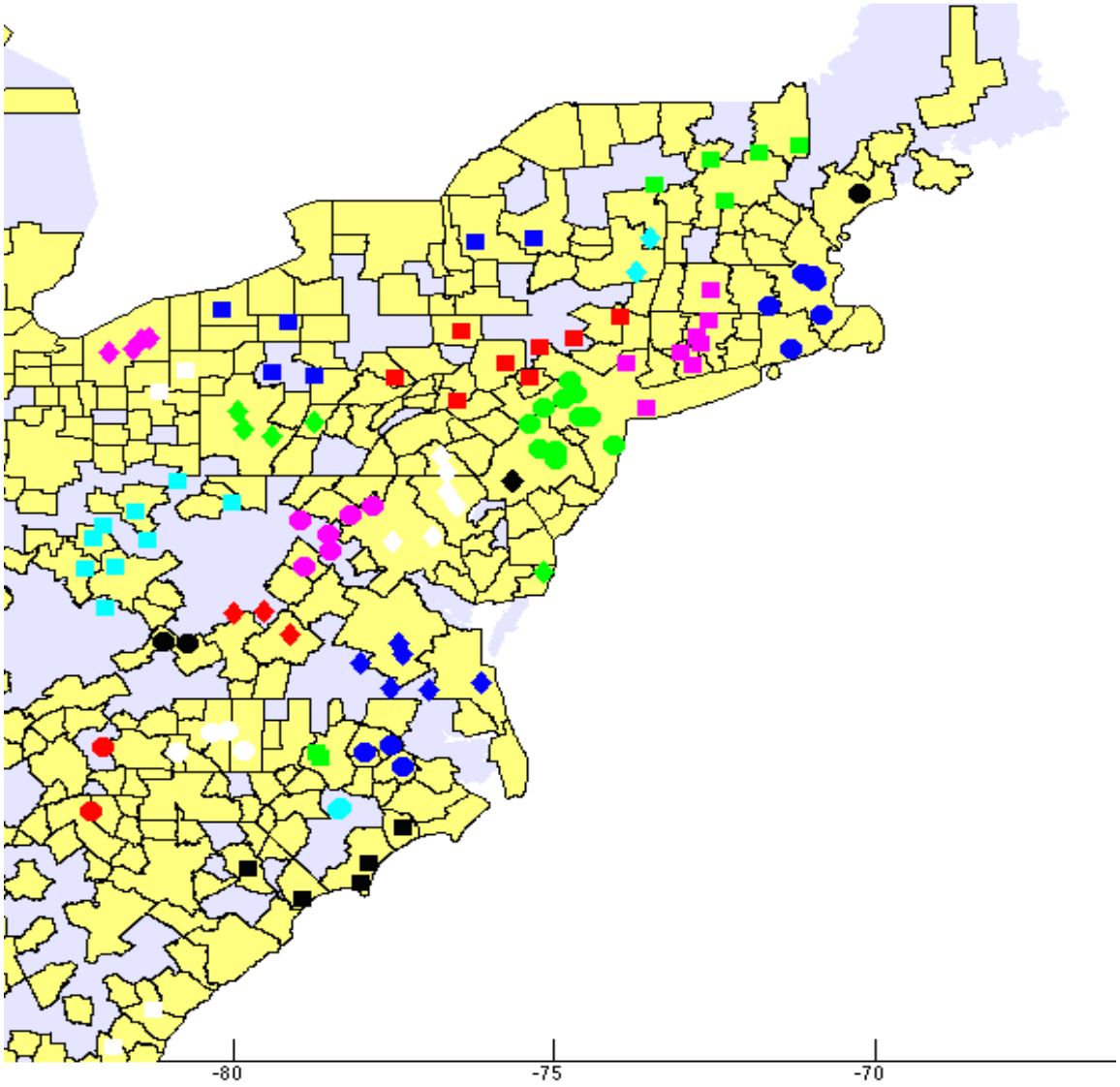

Figure 2.1: *Markets by clustering and CBSAs, 1997Q3*

Table 2.3: *Clustering results comparison with CBSA markets*

	Clustering	CBSA
Max loss per location		
$max_l \left[\lvert S(l) - S(l \in r^*)\rvert / S(l)\right]$	0.005	0.530
Total loss		
$\left[\sum_l \lvert S(l) - S(l \in r^*)\rvert\right] / \left[\sum_l S(l)\right]$	6×10^{-5}	8×10^{-3}
Affected locations		
$\sum_l \mathbb{1}_{S(l) \neq S(l \mid l \in r^*)}$	55	123
Number of locations	241	241

2.2.3 Cost estimation

In this section, I explain the estimation of the cost function. Due to the non-stationary nature of the game, the estimation takes a different approach from the two-stage procedure proposed by BBL or POB. The approach includes two main parts. It first computes the value function using a 'rolling window' approximation. Then the game is solved using backwards induction and the approximated continuation values. The advantage of the approach is that it is more suitable for the counterfactual of interest which is to examine preemptive incentives, and that it is closer to how managers actually make entry decisions, as will be discussed later.

First, I describe the estimation of the parameters outside the dynamic game. Recall firm's static profit of store l is given by (1.3.2). The gross margin, labor cost, and land cost are estimated following Holmes (2011). μ_j is computed using information in firms' annual reports. The average gross margin for Blue and Red firm is 0.24 and 0.22, respectively. The amount of labor and land of a store is assumed to be proportional to revenue:

$$E(R_{jt}^l) = \mu_E R_{jt}^l,$$

$$L(R_{jt}^l) = \mu_L R_{jt}^l.$$

μ_E is calibrated using labor costs listed in firms' annual reports. μ_L is computed using census data and county property tax data of a subset of stores. See Appendix A of Holmes (2011) for details.

The estimated parameters in the dynamic game are per unit distribution cost ψ_j and fixed cost α_j of Blue and Red firm. Recall firms choose the optimal locations given the total number of new stores to open each period by maximizing (1.3.4). Applying two-stage budgeting to this problem with separable markets, firms first choose the optimal locations within each market given the total number of new stores in each market, and then optimize over market level budgets. In the estimation, for computational reasons, I condition on the observed market level budgets and use the information in firms' entry decisions within each market only. In other words, each period, firms solve Equation (1.3.6) for each independent market. Without computational constraints, one can in addition solve the upper level optimization problem to get the optimal number of new stores allocated to each market.

Most of the dynamic game estimation methods in the literature follow a two-stage procedure as in BBL or POB. However, the current problem does not fit in the two-stage estimation framework, for two reasons.

First, in the two-stage framework, the state transition probabilities are estimated non-parametrically in the first stage. This requires observing each state repeatedly in the sample. In the current sample, however, no state is observed more than once. The nature of the problem is non-stationary. Because both firms are expanding in the sample period, the number of stores keep growing and new stores appear every period. Although clustering and market division that are done every period make the evaluation of potential locations focus on local nearby locations and demographics, the set of potential locations are different in each period. To convert the problem into one that fits the two-stage estimation framework, one solution would be to divide different states into groups by some similarity measure and to treat each groups of states as a single state. This is not desirable for a second reason: the state contains rich geographic information that can be important for studying preemptive

incentives. Location characteristics contained in the state variable include the number of competitor's stores, stores belonging to the same firm around them, distance to those stores, distance to the distribution center, and demographic information. They are important information for identifying preemptive incentives. Pooling them into groups may bias the results.

Second, even if the estimation can be done using the two-stage method, the counterfactuals of interest cannot. Using the transition probabilities estimated in the first stage to compute the preemption counterfactual can be problematic. In the counterfactual where preemption is removed, it will be required that players optimize the entry decision without taking into account preemption motives, i.e. firms are not fully optimizing. If the transition probabilities are estimated in the first stage non-parametrically and applied in the counterfactual, it is not guaranteed that those are still the correct transition probabilities when preemption is not allowed. Therefore, the two-stage estimation method is not suitable.

One alternative way to estimate the game is to solve for the nested fixed point as in Pakes (1986) and Rust (1987). In other words, for a fixed parameter value, the dynamic game can be solved and the optimal choice probabilities $Pr(a_{jmt}|s_t)$ can be matched to the observed entry decisions in each market and each period. Then, the above step is repeated for many parameter values to find the estimate that maximizes the likelihood of the observed choice. The difficulty is that instead of the single agent's dynamic problem in Pakes (1986) and Rust (1987), the current problem also has strategic interactions between players, which makes the value function more difficult to compute. Next, I describe how the value function can be approximated using a 'rolling window' method.

Each period, in each market m, firm chooses B_{jmt} locations to open new stores from the L_{mt} potential stores. The set of potential locations L_{mt} is all the locations Blue and Red firm entered between t and the end of the sample period, T. First, the value function can be computed by solving the game using backwards induction. Starting from the last period, T, terminal values can be computed for each possible state s_{mT}. I assume the terminal value equals $\mathbb{E}V(s_{mT})/(1-\beta)$. β is set to be 0.99, i.e. the annual discount factor is 0.95. $\mathbb{E}V(s_{mT})$

can be computed by solving the game for the last two years of data left out of the sample. It allows the decision in period T to be dynamic with respect to the 8 periods ahead, instead of completely static. The implicit assumption is that firms do not foresee any more entry or change in demographics after $T+2$. This is a limitation but no more data is available.

Second, to compute the continuation value for each of the $\binom{L_{mt}}{B_{jmt}}$ states, the value function needs to be computed for each of the possible states in the future between t and $T+2$. This is computationally infeasible. For period $t=1$, $L_t = L = 3123$. Assume clustering can reduce the market size to 30[1], the number of value functions need to be computed is

$$\begin{aligned}
&\binom{L_m}{B_{1m1}} \cdot \binom{L_m - B_{1m1}}{B_{2m1}} \cdots \binom{L_m - \sum_{\tau=1}^{T}\sum_j B_{jm\tau}}{B_{1mT}} \cdot \binom{L_m - \sum_{\tau=t}^{T}\sum_j B_{jm\tau} - B_{1mT}}{B_{2mT}} \\
&\quad = \binom{L_m}{B_{1m}} \cdot \left(\binom{B_{1m}}{B_{1mT}} \cdot \binom{B_{2m}}{B_{2mT}} \cdot \binom{B_{1m} - B_{1mT}}{B_{1m(T-1)}} \cdot \binom{B_{2m} - B_{2mT}}{B_{2m(T-1)}} \cdots \right), \quad (2.2.1)
\end{aligned}$$

where B_{jm} is the total number of new stores that belong to firm j in market m, $j = 1, 2$. Assume half of the stores are Blue, the first term $\binom{30}{15}$ is on the order of 10^8. Moreover, the game has to be solved for each parameter value in the estimation. Therefore, an approximation of the value function is necessary for both the firm and the econometrician.

As discussed above, changing the state variables to reduce the size of the state space is not a desirable approach. One can also reduce the state space by restricting j's choice set to be locations entered by j only. In this case, the first term in (2.2.1) would become 1. However, not allowing firm j to choose from $-j$'s observed locations rules out some of the strategic interactions between the two firms, including preemptive incentives. As an alternative, I restrict the choice set of the firm in each period using a rolling window. For each period t, instead of choosing from all observed locations between t and $T+2$, firms choose from those entered by either firm between t and $t+8$. In other words, j solves the

[1]This is already unrealistic. It implies the locations need to be divided into 100 markets, therefore the loss from clustering must be substantial.

following equation in period t for each market m,

$$V(s_{jmt}, s_{-jmt}) = \max_{a_{jmt} \in \mathbb{A}_{mt}} \left\{ \sum_{\tau=t}^{t+8} \frac{\beta^{\tau-t}}{(1-\beta)^{\mathbb{1}\{\tau=t+8\}}} \sum_{s_{-jm\tau}} \sum_{s_{jm\tau}} \mathbb{E}V(s_{jm\tau}, s_{-jm\tau}) \right.$$
$$\left. P(s_{jm\tau}|s_{jmt} + a_{jmt}, s_{-jmt}) P(s_{-jm\tau}|s_{jmt} + a_{jmt}, s_{-jmt}) \right\} \quad (2.2.2)$$

s.t.

$$\sum_{l=1}^{L_{mt}} a_{jmt}^{l} \leq B_{jmt},$$

where $\bar{L}_t$ is the set of locations entered by Blue and Red firm between t and $t+8$, $(1-\beta)^{\mathbb{1}\{\tau=t+8\}}$ is a scaling factor for terminal period $t+8$. This reduces the computational burden dramatically. Clustering is now done over location $\bar{L}_t$ which is all potential locations between t and $t+8$. The maximum size of market, $\bar{L}_{mt}$ is 12 after taking out a few city centers.

The approximation is close to how managers actually make decisions. According to the managers I interviewed, firms usually have a fairly good idea about how many stores they are planning to open and where the potential locations are in the next two years, which corresponds to 8 periods in my model. Beyond the two year window, it is difficult for managers to foresee how many new markets they are likely to enter or where the desirable locations could be.

However, the approximation imposes two restrictions on firm's optimization problem. First, since the optimization stops at $t+8$, any change of the state variable after $t+8$, and thus any change of the continuation value, is not taken into account. The number of possible paths that need to be evaluated becomes

$$\binom{L_{mt}}{B_{1mt}} \cdot \binom{L_{mt} - B_{1mt}}{B_{2mt}} \cdots \binom{L_{mt} - \sum_{\tau=t}^{t+7} \sum_j B_{jm\tau}}{B_{1mt+8}} \cdot \binom{L_{mt} - \sum_{\tau=t}^{t+7} \sum_j B_{jm\tau} - B_{1mt+8}}{B_{2mt+8}}. \quad (2.2.3)$$

Second, the possible paths firm j is optimizing over between t and $t+8$ are further restricted due to the restriction of choice sets by the rolling window. All the L_t in (2.2.3) becomes $\bar{L}_t$, and therefore the last term in (2.2.3) becomes 1. The computational burden is further reduced. However, the possible paths between t and $t+8$ are restricted to include only

observed locations $\bar{L}_t$, that is, any locations entered after period $t+8$ are not considered as a potential location in period t. I am currently working on testing how sensitive the results are to this restriction by allowing the length of the rolling window to vary.

Finally, I do a grid search through the parameter space and use maximum likelihood to estimate the parameters $\{\psi_j, \alpha_j\}$, $j = 1, 2$. Assume the cost shock of each action follows a type I extreme value distribution. Then the choice probabilities $P(s_{jmt+1}|s_{jmt}, s_{-jmt})$ have a closed form solution. For a given parameter value, I solve the game in (2.2.2) for each market, period, and firm, and match the choice probabilities to the observed choice by:

$$max_{\psi_j,\alpha_j} \sum_m \sum_t \sum_j log\left(Pr(a_{jmt}|s_{jmt}, s_{-jmt}, \psi_j, \alpha_j)^{Y(a_{jmt})}\right), \tag{2.2.4}$$

where

$$Pr(a_{jmt}|s_{jmt}, s_{-jmt}, \psi_j, \alpha_j) = \frac{exp(\mathbb{E}V(s_{jmt} + a_{jmt}, s_{-jmt}))}{\sum_{a_{jmt}\in\mathbb{A}_{jmt}} exp(\mathbb{E}V(s_{jmt} + a_{jmt}, s_{-jmt}))},$$

and $Y(a_{jmt})$ is the indicator of the observed choice, i.e.

$$Y(a_{jmt}) = \begin{cases} 1 & a_{jmt} = a^{observed} \\ 0 & otherwise \end{cases}.$$

2.2.4 Estimation results and interpretation

Table 2.4 presents the estimation results of distribution cost and fixed cost by firm. The distribution cost per thousand-mile is 1.61 million dollars for Blue firm and 0.68 million dollars for Red firm. Using industry sources, Holmes (2011) estimated Blue firm's trucking cost of distribution to be 0.8 million dollars per thousand-mile. Using his model, he estimated the total distribution cost to be around 3.5 million dollars per thousand-mile. My estimate of 1.61 is in the interval of $[0.8, 3.5]$, and closer to the industry source of trucking cost than Holmes' estimate. Holmes interprets his estimate of distribution cost as economies of scale, since it measures the average cost saving of locating a store 1000 miles closer to a distribution center in a single agent's optimization problem. The smaller economies of scale in my results is mainly due to the fact that the model takes into account interactions with Red firm. Since distribution centers are located in rural areas, moving a store closer to a

Table 2.4: *Distribution and fixed cost estimates*

Parameter estimates and 95% confidence intervals	
Blue firm's distribution cost ($1000/mi)	1.61
	[1.23, 1.98]
Red firm's distribution cost ($1000/mi)	0.68
	[0.02, 0.80]
Blue firm's fixed cost ($M)	0.43
	[0.27, 0.55]
Red firm's fixed cost ($M)	0.15
	[0.09, 0.31]
Number of observations	1226
likelihood ratio index	0.32

s.e. are computed using bootstrap and does not include the errors from first stage demand estimation or second stage clustering.

distribution center implies moving away from urban markets where demand is high. This could lead to giving up profitable locations to the competitor, especially, as will be shown below, if the competitor (Red firm) has an urban advantage. Therefore, taking into account the strategic interactions with Red firm, the overall economies of scale is smaller in this model.

The estimates also indicate that per unit distribution cost is lower for Red firm than for Blue firm. However, as shown in Table 1.2, Red stores are on average further away from their own distribution centers than Blue stores are from theirs. For example, the average per store distribution cost in 1990 is 0.22 million dollars for Blue firm and 0.15 million dollars for Red firm. The ratio 0.22/0.15 equals 0.68 which is the ratio of average sizes of Blue stores

and Red stores. Thus the average per store distribution cost conditional on the size of the store is about the same for Blue and Red firm. Moreover, the average per store distribution cost translates to about 0.5% of sales for Blue firm and 1% for Red firm. Therefore, Blue firm has a more efficient distribution system than Red firm, which is consistent with findings by Bradley *et al.* (2002).

The average fixed cost of operating increases by 0.43 million dollars and 0.15 million dollars per year for Blue and Red firm respectively, when population density increases from 250 (25th percentile) to 700 (50th percentile) thousands people per 5-mile radius circular area in 1990. The average fixed cost per store in 1990 is about 1.84 and 0.62 million dollars, or 4% or 5% of sales, for Blue and Red firm, respectively. Since $0.62/1.84 < 0.68$, Red firm has an urban advantage relative to Blue firm. This is consistent with the store characteristics comparisons in Table 1.2.

Standard errors are computed using bootstrap over markets. They do not include estimation errors of demand in the first stage or clustering errors in the second stage. I am currently working on including those errors in the standard errors of the structural estimates using a simulation method. See Appendix II for details of the simulation method.

2.3 Counterfactual I: Preemptive entry

In this section, I conduct a counterfactual analysis to examine the impact of preemptive motives on discount retailers' entry decisions. I remove preemptive motives using a one-period deviation method and compare firms' optimal response to that of the equilibrium outcome with preemption. In other words, what the optimal entry decisions would be if the firm did not need to preempt. I find that preemptive incentives are important to firms' entry decisions and that they lead to an average loss of production efficiency of about 1 million dollars per store, measured by combined sum of current and future profits of the two firms.

It is hard to identify preemptive incentives because they arise in a complex dynamic setting. For preemptive motives to arise, both dynamic optimization and strategic interactions between firms have to be allowed. In such settings, for example, firm j optimizes over

three sets of variables: current state (s_{jt}, s_{-jt}) which I refer to as 'static competition', $\{s_{j\tau}\}_{\tau>t}$ which is the possible state of j and where economies of scale comes from, and $\{s_{-j\tau}\}_{\tau>t}$ which is the opponent's possible state in the future and where preemption comes from. Moreover, as preemption is a motive for acting rather than an action, it cannot be directly observed by the econometrician. Given the complex setting in which preemption arises and its unobserved nature, it is often hard to separate it from static competition between players, incentives to optimize dynamically or unobserved market characteristics. Furthermore, the evaluation of efficiency loss requires that the game setting in the counterfactual not drastically differ from the original setting such that payoffs are comparable under the two settings. Thus solving a different game in which preemption incentives are removed would not be appropriate for the purpose of this counterfactual analysis.

Next, I introduce a formal definition of preemption and present a one-period deviation method that separates preemptive motives from static competition and leads to simple payoff comparisons. I define the preemptive incentives of firm j entering a location l to be the change in $Pr(a^l_{jt}|s_t)$ in response to $Pr(a^l_{-jt'}|s_t)$ in equilibrium, where $Pr(a^l_{jt}|s_t)$ is the choice probability of firm j entering location l at state s_t in equilibrium, and $Pr(a^l_{-jt'}|s_t)$ is the same probability for $-j$ in period t', $t' > t$. In other words, preemptive incentives measure how much, in equilibrium, firm j's likelihood of entering location l today is impacted by its opponent's likelihood of entering the same location in the future, holding static profits constant.

The one period deviation method attempts to measure the change in $Pr(a^l_{jt}|s_t)$ when $Pr(a^l_{-jt'}|s_t)$ is set to zero. The idea is the following: for each of Blue's observed choices, remove those choices from Red firm's choice set for one period. Thus Blue firm knows Red would not be allowed to enter those locations for one period. Then I investigate if Blue firm has profitable deviations by delaying entry at those locations. Specifically, I solve the

following equation to compute the choice probabilities of Blue without preemption,

$$V(s_{jmt}, s_{-jmt}) = \max_{a_{jmt} \in \mathbb{A}_{mt}} \Bigg\{ \sum_{\tau=t}^{t+8} \beta^{\tau-t} \sum_{s_{-jm\tau}} \sum_{s_{jm\tau}} \mathbb{E}V(s_{jm\tau}, s_{-jm\tau}) P(s_{jm\tau} | s_{jmt} + a_{jmt}, s_{-jmt}) \\ P(s_{-jm\tau} | s_{jmt} + a_{jmt}, s_{-jmt}) | a_{-jmt} \neq a^{obs}_{jmt+1} \Bigg\} \quad (2.3.1)$$

s.t.

$$\sum_{l=1}^{L_{mt}} a^{l}_{jmt} \leq B_{jmt},$$

where a^{obs}_{jmt+1} is Blue's observed choice in period t. Note the market level budget constraint is held fixed, so that Blue firm is not fully optimizing. Restricting $a_{-jmt} \neq a^{obs}_{jmt+1}$ gives Blue firm an advantage over the set of locations in a_{-jmt} and the firm's payoff is at least as high as in the original equilibrium. Thus relaxing the market budget $\bar{L}_{mt}$ would only lead to higher payoff. As a result, if Blue firm's payoff increases by solving Equation (2.3.1), the amount of payoff increase is a lower bound, i.e. the impact of preemptive incentives shown by the above procedure is a lower bound.

For the opponent, Red firm, the optimization problem is the same as in (2.3.1) with j and $-j$ switched. If Blue firm's choice probabilities stay the same, Red firm would also stay in the original equilibrium. If Blue firm deviates, Red firm is allowed to respond. Note that in this case, the market budget $\bar{L}_{mt}$ is also held fixed for Red firm. This does not bias the result. Red firm is deprived by being forced to choose from a smaller set of locations for one period. If Blue firm were to deviate and delay entry, and Red firm were fully optimizing, it would switch away to other markets, inducing a weaker presence in the current market. This would lead to higher payoff for Blue firm. Therefore the impact of preemptive incentives measured in this experiment is a lower bound.

Results are presented in Table 2.5. I compute the choice probabilities of Blue firm entering the observed locations when preemptive incentives are removed for one period. For cases in which it is profitable for Blue firm to deviate from the original equilibrium, I compute the payoff increase from the original equilibrium payoff. Out of 1278 locations, there are 425 locations at which preemption is observed in this experiment. There is profitable deviation

Table 2.5: *Preemption: one period deviation of Blue firm*

Choice probabilities and payoffs	
Average payoff increase($M)	0.86
Average choice probability decrease	0.14
Number of delayed entries	425
Total number of obs.	1278
Number of Blue stores out of 425	392
Efficiency loss ($M)	397

for Blue firm to delay entry when those choices are taken out of Red's choice set for one period. For those 425 locations where preemption is observed, average choice probability decreases by 0.14 compared to the choice probabilities in the original equilibrium. Blue firm's payoff could increase by an average of 0.86 million dollars for each delayed entry, which is about a small store's one year profits.

Then I study loss production efficiency due to preemption. Since I do not have enough data to conduct total welfare analysis, in the current counterfactual, I compute efficiency loss from firms' perspectives only, by examining the change of sum of expected current and future profits of both firms. Results are shown in the last two rows of Table 2.5. Out of the 425 locations, there are 392 locations where the expected total payoff of Blue and Red firm is higher in the counterfactual than in the original equilibrium. These are the locations at which the higher payoff of Blue firm in the counterfactual is enough to compensate the lower payoff of Red firm due to its restricted choice set in one period. The total amount of payoff increase is almost 397 million dollars. On average, the efficiency loss per location is 1.01 million dollars, which is about the annual profit of a small-to-median size store. For the reasons discussed above, it is a lower bound of the efficiency loss in consequence of preemption. The market level budgets are held fixed and firms are not fully optimizing.

Table 2.6: *Preemption vs. no preemption: location comparison*

Characteristics of preemption and no preemption locations		
	Preemption	No preemption
Distance to Blue DC (mi)	217.94	208.50
Distance to Red DC (mi)	273.22	287.35
Total Store density	24.62	22.56
Blue store density	12.14	11.55
Population density (1000)	175.48	172.12
Number of observations	425	853

From the perspective of the production side, preemption results in a substantial loss of production efficiency.

Next I investigate why evidence of preemption is found at 425 locations but not at others. One reason could be that preemption may not always be profitable. I compare the two sets of locations in Table 2.6 and refer to them as preemption and non-preemption locations. First of all, distance to Blue's distribution center is smaller for the non-preemption stores than for the preemption stores. The opposite is true for distance to Red's distribution centers. This is consistent with the motives of preemption in the current experiment. Since preemption is more profitable at locations where the opponent is more likely to enter in the future, those locations are likely to be closer to Red distribution centers. On the other hand, delaying entry at these locations would mean giving up current profits, hence it is more likely to observe preemption in the current experiment at locations where current profits are low. This is consistent with longer distance to Blue's distribution center. Therefore, in the current one-period deviation experiment, it is more likely to observe preemption at locations that are closer to Red distribution center and further away from Blue distribution center. Moreover, preemption stores also tend to locate in areas with higher store density

Table 2.7: *Preemption: response of Red firm if Blue did not enter*

Change of choice probabilities and payoff	
Average payoff increase($M)	2.99
Probability of entry	0.77
Number of entries	472
Total number of obs.	1278

and higher population density. This is also consistent with the fact that those are the areas that Red store is more likely to enter.

Finally, I compute the one-period deviation for Red firm. In this experiment, Blue firm is assumed to stay away from its observed choices for one period and Red firm is given the opportunity of choosing from those locations in the next period. I examine if Red firm would choose to enter those locations, and if so how much its payoff would increase. In this case, Red firm's budget constraint for each market is allowed to adjust, since in the original equilibrium Red firm may not have entered any location in the same market. However, this is not computationally burdensome since each market is evaluated independently in this experiment. For example, when evaluating Red firm's deviation in market m, all the other markets $-m$ are held fixed at the original equilibrium. Thus, to determine if it is profitable, the payoff of deviation is compared to the median expected payoff of $-m$ markets.

The results are presented in Table 2.7. There is about one third of the locations where it is profitable for Red firm to enter right away, had Blue firm stayed out of those locations. The average probability of entry is 0.77. The average payoff increase is 2.99 million dollars for each location, which is about a big store's one year profits.

In the current literature, there are two methods of identifying preemption using empirical analysis. The first method, used in Schmidt-Dengler (2006), separates preemption by solving a pre-commitment game following the theoretical work of Reinganum (1981). In the pre-commitment game, firms make the entry decision in the first period for the following T

periods and commit to it. The problem of this approach is two-fold. First, it does not exclude preemption completely. It prohibits a firm from responding instantaneously to the opponent's action, but the firm is still optimizing in period 1 taking into account the possible actions of the opponent in the future. Second, since pre-commitment games demand a different setting and lead to completely different equilibria, it is difficult to compare payoffs with the ones in the original game.

The second method is to solve a single agent's dynamic problem, as in Igami and Yang (2014). Applying their method in the current setting, Red firm would make entry decisions assuming Blue firm would not enter any market in any period. From Blue firm's perspective, Red firm has become a part of nature and does not respond to Blue's actions, and therefore cannot be preempted. Thus Blue firm solves a single agent's dynamic optimization problem. This experiment does not properly separate preemptive incentives since it also precludes Red firm from responding to Blue firm's actions in the current period, i.e. it prohibits static competition. Setting preemptive motives motives apart, it is not clear why, when Blue's entry is observed, Red firm should optimize assuming Blue firm is not entering any market.

2.4 Counterfactual II: Subsidy Policy after Red Firm Exits

Red firm started exiting in many markets by closing stores in 2000. It has closed more than 1000 stores in the past 15 years. The store closings have a big impact on local economies (Shoag and Veuger, 2014). Local governments have proposed to subsidize Red firm to stay or other retailers such as Blue firm to enter. For example, Buffalo, NY proposed a 400,000 dollar subsidy for Red firm to stay[2]. Rolling Meadows, IL managed to subsidize Blue firm to enter after their Red store closed[3]. However, lots of ex-Red retail slots remained empty years after Red firm's exit. The example of Rockledge, FL I have mentioned. Indiana Harbor

[2]Source: `http://www.huffingtonpost.com/2012/01/26/sears-closes-cities_n_1231326.html`

[3]Source: Good Jobs First and Corporate Research Project

Beach, FL is another example[4]. The retail space of the former Red store stayed empty for 12 years. Therefore, it seems important for policy makers to better understand the impact of government subsidies on retailer's entry decisions. The current entry model can be used to answer this question. It is also important to consider the welfare loss due to the closing of stores. Although I do not have enough data to conduct total welfare analysis, I compute the loss of consumer drive time due to store closings and compare it to the size of the observed subsidies. This section attempts to answer those two questions by computing payoff differences between ex-Red locations and the rest of potential Blue firm's locations and deriving welfare loss to consumers caused by higher travel cost of shopping.

First, I compute the expected payoff of Blue firm entering each ex-Red locations and compare it to the expected payoff of Blue firm entering each of the other potential locations for each period between year 2000 and 2003. There are 96 Red store closings in those four years and the number of potential locations is 815. I refer to the difference between the payoff of the median store in the two groups as the 'subsidy', since it is the amount of extra payoff needed for the ex-Red locations to become as profitable as the other potential locations. Then I compute subsidies separately for two sub-periods: 2000-2001, and 2002-2003. The difference between those two sub-periods is that between 2000 and 2001, Red firm was still expanding, but beginning in 2002 the expansion stopped. This makes a difference for Blue firm's incentives to enter ex-Red locations. In the first sub period, it is clear that Blue firm knew that Red firm would not re-enter at ex-Red locations after the store closings, which removes the preemptive motives for Blue to enter those locations, compared with other potential locations. On the other hand, in the second sub-period, there is no preemptive motive for Blue firm at any potential location including the ex-Red ones since the expansion of Red firm has stopped. Thus it does not make the ex-Red locations less favorable, as in the first sub-period.

Results are reported in row 1 of Table 2.8. The median size of subsidies in the period

[4]Source: `http://www.floridatoday.com/story/money/business/2014/07/27/kmart-goes-next/13197001/`

Table 2.8: *Subsidies before and after Red firm stops expanding*

Average subsidies per store

	Total	2000-2001	2002-2003
Median predicted subsidy ($M)	3.21	5.13	1.29
Number of subsidies $\leq$ 0.5M per period	5.5	3.5	7.5

between 2000 and 2003 is 3.21 million dollars per store. The same measure is higher for period 2000 and 2001 when Red firm is still expanding: 5.13 million dollars per store. The stop of Red firm's expansion makes ex-Red locations less unattractive compared to other potential locations, with a median subsidy amount of 1.29 million dollars per store. The difference between the subsidies needed for Blue to enter in the two sub-periods gives one explanation to why some retail slots remained empty for years after a store closing. The preemptive motives lead to the unattractiveness of those locations compared to other potential locations.

To better interpret the sizes of the subsidies, I compare them to the size of observed subsidies given to Blue firm between 2000 and 2014. The subsidy data is obtained from goodjobsfirst.org. Although the list of subsidies is incomplete and some of the subsidy sizes are approximated, it gives a general sense of the size of the subsidies. The average size is about 0.5 million dollars. Accordingly, I count the number of ex-Red locations whose payoffs are less than half a million dollar lower than the payoff of the median potential location that has never been entered by either firm. Row 2 of Table 2.8 reports the results. On average, there are 5.5 ex-Red locations per period that Blue firm would enter with a subsidy of 0.5 million dollars. The number drops to 3.5 for the period of 2000-2001, while an increase of 4 locations period is observed for the period of 2002-2003. Overall, the observed average subsidy does not have a big impact on Blue firm's entry at locations where Red firm exited.

Lastly, I examine the welfare loss of consumers due to increased travel time to discount

Table 2.9: *Consumer welfare loss due to store closings*

Consumer drive time loss per store per year

	Total	2000-2001	2002-2003
Distance per person (mi)	4.05	4.10	3.99
Total distance (10^5mi)	8.70	9.80	7.60
Total welfare loss ($M)	1.26	1.42	1.10

shops when Red stores close. This is not expected to measure the total welfare change due to store closings. Other factors such as employment, impact on small businesses, local government income are also affected by exits of big box retail stores (Basker, 2007,Jia, 2008). However, this analysis still helps to get a sense of whether, in general, the welfare loss is comparable to the size of subsidies. I use the demand model to compute the change of distance between a consumer and a store in the consumer's choice set due to each of Red store's closings. Note some consumers may switch to the outside option after a Red store closes. For those consumers, I assume the distance traveled to the outside option to be the average distance travelled by a consumer to a discount retail store, 15 miles. This measure comes from the industry survey data collected by Fox *et al.* (2004).

Table 2.9 reports the results. The average travel distance per person increases by 4.05 miles between 2000 and 2003, while the total distance increases by 870 thousands miles. The total welfare loss per year is computed using the following formula: total distance/40mph$\times$7.25(federal min. wage)$\times$10 trips$\times$2(round trip)/2.5(avg. household size). Total distance divided by driving speed of 40 mile per hour is the total time of travel which is multiplied by the federal minimum wage to get the dollar value. I assume a consumer makes 10 trips per year to a discount store. Given the estimated annual spending is $2444, it seems reasonable to assume she spends about $240 on each trip. 10 trips per year is also much lower than the estimate by Fox *et al.* (2004) of visiting a discount store every two weeks, which squares with the goal of finding a lower bound for consumer welfare

loss. Then the number is multiplied by 2 to account for round trips and divided by the average household size from census data, assuming one person shops for each household. The total welfare loss amounts to 1.26 million dollars. Although it is much smaller than the average subsidy size of 3.21 million dollars, the break down of 1.10 million dollars for period 2002-2003 is very close to the 1.29 million dollar subsidy for this period. Therefore, welfare loss to consumers due to store closings can be substantial and can make the subsidy on entry worthwhile.

2.5 Conclusion

This paper studies how multi-store retail chains make entry decisions, with a special emphasis on the impact of preemptive incentives.

The study is carried out in a dynamic duopoly model in which firms make entry decisions at spatially interdependent locations. It is shown that the model can be made tractable by applying two-stage budgeting and separability. Instead of using census geographic units, market divisions are inferred using machine learning tools built on separability conditions, so that the spatial interdependence across store locations is preserved. The estimation is carried out by solving the game using backwards induction and applying a 'rolling window' approximation to compute the value function. This model and this estimation method can be applied to other retail industries or sectors in which network effects or cost sharing are present. More generally, the application of machine learning tools in structural estimation and its impact on inference is an interesting direction for future research.

Counterfactual analyses are also conducted. The results suggest that preemptive incentives are important in multi-store retailers entry decisions and that they can lead to substantial efficiency loss. When a retailer exits a market, as frequently observed in the recent crisis, the store location becomes less attractive to other retailers due to the absence of preemptive incentives. In these cases, although consumer welfare loss from the store closings can be significant, standard government subsidies prove insufficient to encourage entry. The framework presented here can be used to assess other public policy issues that

arise in those industries the model can be applied to.

Chapter 3

Inference for Misspecified Models with Fixed Regressors [1]

3.1 Introduction

Following the seminal work by Eicker (1967), Huber (1967) and White (1980ab, 1982), researchers estimating regression functions routinely report standard errors that are robust to misspecification of the models that are being estimated. Müller (2013) gave the corresponding confidence intervals a Bayesian interpretation. A key feature of the approach developed by Eicker, Huber, and White (EHW from hereon), is that in regression settings it focusses on the best linear predictor that minimizes the distance between a linear function and the true conditional expectation, averaged over the joint distribution of all variables, with a similar interpretation in nonlinear settings. We argue that in some regression settings it may be more appropriate to focus on the conditional best linear predictor defined by minimizing this distance averaged over the empirical instead of the population distribution of the covariates. The first contribution of this paper is to extend the EHW results to such settings. For a large class of estimators, including maximum likelihood and method of moment estimators, we first formally characterize the generalization to nonlinear models of the conditional best

[1] Co-authored with Alberto Abadie and Guido Imbens

linear predictor. We then derive a large sample approximation to the variance of the least squares and method of moments estimators relative to this conditional estimand. In general, in misspecified models, the robust variance for the conditional estimand is smaller than or equal to the EHW robust variance. Second, we propose a consistent estimator for this variance so that asymptotically valid confidence intervals can be constructed. The proposed estimator generalizes the variance estimator proposed by Abadie and Imbens (2006) for matching estimators and more generally the differencing methods used in Yatchew (1997, 1999). In correctly specified models the new variance estimator is simply an alternative to the standard EHW robust variance estimator. In misspecified models it is the only consistent estimator available for the asymptotic variance for the estimand conditional on covariates.

Whether conditional or unconditional estimand should be the primary focus is context specific and we do not take the position that either the conditional or unconditional estimand is always the appropriate one. We discuss some examples, first to clarify the distinctions between the two estimands and, second, to make an argument for our view that in some settings the conditional estimand, corresponding to the fixed regressor notion, is of interest. For example, we argue that in cases where the sample is the population there is a strong case for using the estimand conditional on at least some covariates, see also Abadie *et al.* (2014). Such cases are common in economic analyses, *e.g.*, when analysing data where the units are all states of the United States, or all countries of the world. Most importantly, we argue that there is a choice to be made by the researcher that has direct implications for inference. In making this choice the researcher should bear in mind that the variance for the conditional estimand is generally smaller than that for the population or unconditional estimand, and thus tests for the former will generally have better power than tests for the latter.

Note that although we focus on estimands defined in terms of the finite sample distribution of the covariates, our inference relies on large sample approximations. To focus on the conceptual contribution of the current paper and maintain comparability with the preceeding literature, we focus on unconditional inference.

The rest of this paper is organized as follows. Section 3.2 contains a heuristic discussion of the conceptual issues raised by this article in a linear regression model setting. In Section 3.3 we discuss the motivation for the conditional estimand. Next, in Section 3.4 we present formal results covering least squares, maximum likelihood, and method of moments estimators. In Section 3.5 we apply the methods developed in this paper to a data set previously analyzed by Sachs and Warner (1997) to study the relation between country-level growth rates and government fiscal policies. In Section 3.6 we present two simulation studies, one in a linear and one in a nonlinear setting. Section 3.7 concludes. The appendix contains proofs.

Some people just cite papers in introductions for no reason. Anderson and Rubin (1949); Pearson (1901); Spearman (1904).

3.2 The Conditional Best Linear Predictor

In this section we lay out some of the conceptual issues in this paper informally in the setting of a linear regression model. In Section 3.4 we provide formal results, covering both this linear model setting and more general cases including maximum likelihood and method of moments.

Consider the standard linear model

$$Y_i = X_i'\theta + \varepsilon_i, \tag{3.2.1}$$

with Y_i the outcome of interest, X_i a K-vector of observed covariates, possibly including an intercept, and ε_i an unobserved error. Let $\mathbf{X}$, $\mathbf{Y}$, and $\boldsymbol{\varepsilon}$ be the $N \times K$ matrix with ith row equal to X_i', the N-vector with ith element equal to Y_i, and the N-vector with ith element equal to ε_i, respectively. In this setting researchers have often assumed homoskedasticity, independence of the errors terms, and Normality of the error terms,

$$\boldsymbol{\varepsilon}|\mathbf{X} \sim \mathcal{N}(0, \sigma^2 \cdot I_N),$$

where I_N is the $N \times N$ identity matrix. Under those assumptions the exact (conditional)

distribution of the least squares estimator

$$\hat{\theta} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\left(\mathbf{X}'\mathbf{Y}\right),$$

is Normal:

$$\hat{\theta}|\mathbf{X} \sim \mathcal{N}\left(\theta, \sigma^2 \cdot \left(\mathbf{X}'\mathbf{X}\right)^{-1}\right).$$

However, assumptions of linearity of the regression function, independence, homoskedasticity, and Normality of the error terms are often unrealistic. Eicker (1967), Huber (1967), and White (1980ab), considered the properties of the least squares estimator $\hat{\theta}$ under substantially weaker assumptions. For the most general case one needs to define the estimand if the regression function is not linear. Suppose the sample $(Y_i, X_i)_{i=1}^N$ is a random sample from a large population satisfying some moment restrictions. Let $\mu(x) = \mathbb{E}[Y_i|X_i = x]$ be the conditional expectation of Y_i given $X_i = x$, and let $\sigma^2(x)$ be the conditional variance. Even if this conditional expectation $\mu(x)$ is not linear, one might still wish to approximate it by a linear function $x'\theta$, and be interested in the value of the slope coefficient of this linear approximation. Traditionally the optimal approximation is defined as the value of θ that minimizes the expectation of the squared difference between the outcomes and the linear approximation to the regression function. This is generally referred to as the *best linear predictor*, formally defined as

$$\theta_{\text{pop}} = \arg\min_{\theta} \mathbb{E}\left[\left(Y_i - X_i'\theta\right)^2\right]. \tag{3.2.2}$$

Because

$$\mathbb{E}\left[\left(Y_i - X_i'\theta\right)^2\right] = \mathbb{E}\left[\left(\mu(X_i) - X_i'\theta\right)^2\right] + E\left[\sigma^2(X_i)\right],$$

with the last term free of dependence on θ, it follows that we can characterize θ_{pop} as

$$\theta_{\text{pop}} = \arg\min_{\theta} \mathbb{E}\left[\left(\mu(X_i) - X_i'\theta\right)^2\right] = \left(\mathbb{E}\left[X_i X_i'\right]\right)^{-1}\left(\mathbb{E}\left[X_i \mu(X_i)\right]\right),$$

which in turn shows that θ_{pop} can be interpreted as the value of θ that minimizes the discrepancy between the true regression function $\mu(x)$ and the linear approximation, weighted by the population distribution of the covariates.

The results in EHW imply that, under some regularity conditions,

$$\sqrt{N} \cdot \left(\hat{\theta} - \theta_{\text{pop}}\right) \xrightarrow{d} \mathcal{N}\left(0, \mathbb{V}_{\text{pop}}\right),$$

where the asymptotic variance is

$$\mathbb{V}_{\text{pop}} = \left(\mathbb{E}\left[X_i X_i'\right]\right)^{-1} \left(\mathbb{E}\left[(Y_i - X_i'\theta_{\text{pop}})^2 X_i X_i'\right]\right) \left(\mathbb{E}\left[X_i X_i'\right]\right)^{-1}. \tag{3.2.3}$$

White also proposed a consistent estimator for $\mathbb{V}_{\text{pop}}$,

$$\hat{\mathbb{V}}_{\text{pop}} = \left(\frac{1}{N}\sum_{i=1}^{N} X_i X_i'\right)^{-1} \left(\frac{1}{N}\sum_{i=1}^{N} (Y_i - X_i'\hat{\theta})^2 X_i X_i'\right) \left(\frac{1}{N}\sum_{i=1}^{N} X_i X_i'\right)^{-1}. \tag{3.2.4}$$

Using the EHW variance estimator $\hat{\mathbb{V}}_{\text{pop}}$ is currently the standard practice in empirical work in economics. See Angrist and Pischke (2008) for an example and Imbens and Kolesar (2012) for a discussion for finite sample improvements. Resampling methods such as the jackknife and the bootstrap (Efron, 1982; Efron and Tibshirani, 1994) can also be used to construct confidence intervals for θ_{pop}.

In this paper we explore an alternative linear approximation to the possibly nonlinear regression function $\mu(x)$. Instead of minimizing the marginal expectation of the squared difference between the outcomes and the regression function, we minimize this expectation conditional on the observed covariates. Define the *conditional best linear predictor* $\theta_{\text{cond}}(\mathbf{X})$ as

$$\theta_{\text{cond}}(\mathbf{X}) = \arg\min_{\theta} \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}\left[\left(Y_i - X_i'\theta\right)^2 \middle| \mathbf{X}\right]. \tag{3.2.5}$$

The difference with the best linear predictor defined in (3.2.2) is that in (3.2.5) the expectation is taken over the empirical distribution of the covariates, whereas in (3.2.2) the expectation is taken over the population distribution of the covariates. To be explicit about the dependence of the conditional best linear predictor on the sample values of the covariates we write $\theta_{\text{cond}}(\mathbf{X})$ as a function of the matrix of covariate values $\mathbf{X}$. Denoting the N-vector with i-th element equal to $\mu(X_i)$ by $\boldsymbol{\mu}(\mathbf{X})$, we can write $\theta_{\text{cond}}(\mathbf{X})$ as

$$\theta_{\text{cond}}(\mathbf{X}) = \arg\min_{\theta} \frac{1}{N}\sum_{i=1}^{N} \left(\mu(X_i) - X_i'\theta\right)^2 = \left(\mathbf{X}'\mathbf{X}\right)^{-1} \left(\mathbf{X}'\boldsymbol{\mu}(\mathbf{X})\right),$$

to stress the interpretation of $\theta_{\text{cond}}(\mathbf{X})$ as the best approximation to the true regression function, now with the weights based on the empirical distribution of the covariates. Both θ_{pop} and $\theta_{\text{cond}}(\mathbf{X})$ base the linear approximation to $\mu(x)$ on a minimizing of the squared difference between the true regression function $\mu(x)$ and the linear approximation $x'\theta$. The difference between the two approximations is solely in how they weight, as a function of the covariates, the squared difference between the regression function and the linear approximation for each x. The first approximation, leading to θ_{pop}, uses the population distribution of the covariates. The second approximation, leading to $\theta_{\text{cond}}(\mathbf{X})$, uses the empirical distribution of the covariates.

We defer to Section 3.3 the important question whether, and why, in a specific application, $\theta_{\text{cond}}(\mathbf{X})$ rather than θ_{pop} might be the object of interest. In some applications we argue that θ_{pop} is the estimand of interest. However, as discussed in detail in Section 3.3, we also think that in other applications $\theta_{\text{cond}}(\mathbf{X})$ is of more interest than θ_{pop}. Given that the main focus of the previous literature is on population parameters like θ_{pop}, we view the question of inference for $\theta_{\text{cond}}(\mathbf{X})$ as of general interest.

Next we point out the implications of the difference between θ_{pop} and $\theta_{\text{cond}}(\mathbf{X})$. The first issue to note is that for point estimation it is irrelevant whether we are interested in θ_{pop} or $\theta_{\text{cond}}(\mathbf{X})$. In both cases the least squares estimator $\hat{\theta}$ is the natural estimator. However, for inference it does matter whether we are interested in estimating θ_{pop} or $\theta_{\text{cond}}(\mathbf{X})$, unless $\mathbb{E}[\varepsilon|\mathbf{X}] = 0$ and the conditional expectation is truly linear. Consider the variance of the least squares estimator $\hat{\theta}$, viewed as an estimator of $\theta_{\text{cond}}(\mathbf{X})$. The exact (conditional) variance of $\hat{\theta}$ is

$$\begin{aligned}\mathbb{V}\left(\hat{\theta}\middle|\,\mathbf{X}\right) &= \mathbb{E}\left[\left(\hat{\theta}-\theta_{\text{cond}}(\mathbf{X})\right)\left(\hat{\theta}-\theta_{\text{cond}}(\mathbf{X})\right)'\middle|\,\mathbf{X}\right] \\ &= \frac{1}{N}\left(\mathbf{X}'\mathbf{X}/N\right)^{-1}\left(\frac{1}{N}\sum_{i=1}^{N}\sigma^2(X_i)X_iX_i'\right)\left(\mathbf{X}'\mathbf{X}/N\right)^{-1}.\end{aligned} \tag{3.2.6}$$

Directly comparing the normalized variance $N\cdot\mathbb{V}(\hat{\theta}|\mathbf{X})$ to the EHW variance $\mathbb{V}_{\text{pop}}$ is complicated by the fact that $N\cdot\mathbb{V}(\hat{\theta}|\mathbf{X})$ is a conditional variance, rather than an asymptotic variance like $\mathbb{V}_{\text{pop}}$. We therefore look at the unconditional variance of the ols estimator $\hat{\theta}$

as an estimator of $\theta_{\text{cond}}(\mathbf{X})$. Because $\hat{\theta}$ is unbiased for $\theta_{\text{cond}}(\mathbf{X})$, it follows that the marginal variance is the expected value of the conditional variance. Under random sampling the asymptotic variance is

$$\mathbb{V}_{\text{cond}} = \text{plim}\left(N \cdot \mathbb{V}(\hat{\theta}|\mathbf{X})\right) = \left(\mathbb{E}\left[X_i X_i'\right]\right)^{-1}\left(\mathbb{E}\left[\sigma^2(X_i) X_i X_i'\right]\right)\left(\mathbb{E}\left[X_i X_i'\right]\right)^{-1}, \tag{3.2.7}$$

and we have, under regularity conditions, a large sample approximation to the distribution of $\sqrt{N} \cdot \left(\hat{\theta} - \theta_{\text{cond}}(\mathbf{X})\right)$:

$$\sqrt{N} \cdot \left(\hat{\theta} - \theta_{\text{cond}}(\mathbf{X})\right) \xrightarrow{d} \mathcal{N}\left(0, \mathbb{V}_{\text{cond}}\right).$$

The key difference between the robust variance $\mathbb{V}_{\text{pop}}$ proposed by White and the robust variance $\mathbb{V}_{\text{cond}}$ arises from the difference between the conditional variance $\sigma^2(X_i)$ in (3.2.7) and the expectation of the squared residual $\mathbb{E}[(Y_i - X_i'\theta_{\text{pop}})^2|X_i]$ in (3.2.3). The latter is in general larger:

$$\mathbb{E}[(Y_i - X_i'\theta_{\text{pop}})^2|X_i] = \sigma^2(X_i) + (\mu(X_i) - X_i'\theta_{\text{pop}})^2,$$

where $\mu(X_i) - X_i'\theta_{\text{pop}}$ captures the difference between the linear approximation and the conditional expectation. For the asymptotic variances of $\hat{\theta}$ we have:

$$\mathbb{V}_{\text{pop}} = \mathbb{V}_{\text{cond}} + \mathbb{V}(\theta_{\text{cond}}(\mathbf{X})), \tag{3.2.8}$$

where

$$\mathbb{V}(\theta_{\text{cond}}(\mathbf{X})) = \text{plim}\ N \cdot \mathbb{E}\left[\left(\theta_{\text{cond}}(\mathbf{X}) - \theta_{\text{pop}}\right)\left(\theta_{\text{cond}}(\mathbf{X}) - \theta_{\text{pop}}\right)'\right] \tag{3.2.9}$$

The last expectation is over the distribution of $\theta_{\text{cond}}(\mathbf{X})$ as a function of $\mathbf{X}$. Thus in general $\mathbb{V}_{\text{pop}}$ exceeds $\mathbb{V}_{\text{cond}}$, and as a result inference based on $\mathbb{V}_{\text{pop}}$ is conservative for $\theta_{\text{cond}}(\mathbf{X})$. The difference between the two variances is the result of the misspecification in the regression function, that is, the difference between the conditional expectation and the best linear predictor, $\mu(x) - x'\theta_{\text{pop}}$.

The final question we address in this section is how to estimate $\mathbb{V}_{\text{cond}}$. Simple bootstrapping methods do not work (Tibshirani, 1986; Wu, 1986). The challenge is that the conditional variance function $\sigma^2(x)$ is generally unknown. Estimating this is straightforward in the

case with discrete covariates. One can consistently estimate the conditional variance $\sigma^2(X_i)$ at each distinct value of the covariates and plug that in (3.2.7), followed by replacing the expectations by averages over the sample. If the covariates are continuous, however, this is not feasible. In the remainder of this discussion we focus on the continuous covariate case. Dealing with the setting where some of the covariates are discrete is conceptually straightforward, but would require carrying along additional notation and come at the expense of clarity. In the continuous covariate case estimating $\sigma^2(x)$ consistently for all x would require nonparametric estimation involving bandwidth choices. Such an estimator would be more complicated than the EHW robust variance estimator which simply uses squared residuals to estimate the expectation of the squared errors. Here we build on work by Yatchew (1997, 1999) and Abadie and Imbens (2006, 2008b,a) to develop a general estimator for $\mathbb{V}_{\text{cond}}$ that does not require consistent estimation of $\sigma^2(x)$, much like the EHW variance estimator does not consistently estimate $\mathbb{E}[(Y_i - X_i'\theta_{\text{pop}})^2|X_i = x]$ for all x. Let V_X be the covariance matrix of X, $V_X = \sum_{i=1}^N (X_i - \overline{X})(X_i - \overline{X})'/N$, where $\overline{X} = \sum_{i=1}^N X_i/N$. Next define $\ell_X(i)$ to be the index of the unit closest to i in terms of X:

$$\ell_X(i) = \arg \min_{j \in \{1,\ldots,N\}, j \neq i} \left\| X_i - X_j \right\|, \tag{3.2.10}$$

where the norm we use is the Mahalanobis distance, $\|x\| = x' V_X^{-1} x$, although others could be used. Then our proposed variance estimator is

$$\widehat{\mathbb{V}}_{\text{cond}} = \left(\frac{1}{N} \sum_{i=1}^N X_i X_i' \right)^{-1} \tag{3.2.11}$$

$$\cdot \left(\frac{1}{2N} \sum_{i=1}^N \left(\hat{\varepsilon}_i X_i - \hat{\varepsilon}_{\ell_X(i)} X_{\ell_X(i)} \right) \left(\hat{\varepsilon}_i X_i - \hat{\varepsilon}_{\ell_X(i)} X_{\ell_X(i)} \right)' \right) \cdot \left(\frac{1}{N} \sum_{i=1}^N X_i X_i' \right)^{-1}.$$

In Section 3.4 we show in a more general setting that this variance estimator is consistent for $\mathbb{V}_{\text{cond}}$. An alternative estimator for $\mathbb{V}_{\text{cond}}$ exploits the fact that the conditional variance of $\varepsilon_i X_i$ conditional on X_i is the same as X_i times the conditional variance of ε_i given X_i,

$$\widetilde{\mathbb{V}}_{\text{cond}} = \left(\frac{1}{N} \sum_{i=1}^N X_i X_i' \right)^{-1} \cdot \left(\frac{1}{2N} \sum_{i=1}^N \left(\hat{\varepsilon}_i - \hat{\varepsilon}_{\ell_X(i)} \right)^2 X_i X_i' \right) \cdot \left(\frac{1}{N} \sum_{i=1}^N X_i X_i' \right)^{-1}.$$

Although in this linear regression case with conditioning on all covariates both $\widehat{\mathbb{V}}_{\text{cond}}$ and $\widetilde{\mathbb{V}}_{\text{cond}}$ are consistent for $\mathbb{V}_{\text{cond}}$, for nonlinear settings, or with conditioning on a subset of the covariates, only the first estimator $\widehat{\mathbb{V}}_{\text{cond}}$ generalizes. To be specific, suppose that the covariate vector X_i can be partitioned as $X_i = (X_{1i}', X_{2i})'$ and correspondingly $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$, and suppose we wish to estimate the variance conditional on $\mathbf{X}_1$ only. In this case the probability limit of the normalized variance for the least squares estimator is

$$\mathbb{V}_{\text{cond}} = \left(\mathbb{E}\left[X_i X_i'\right]\right)^{-1} \left(\mathbb{E}\left[\mathbb{V}\left(\varepsilon_i X_i \middle| X_{1i}\right)\right]\right) \left(\mathbb{E}\left[X_i X_i'\right]\right)^{-1}. \tag{3.2.12}$$

Our proposed estimator for this conditional variance is

$$\widehat{\mathbb{V}}_{\text{cond}} = \left(\frac{1}{N}\sum_{i=1}^{N} X_i X_i'\right)^{-1} \tag{3.2.13}$$

$$\cdot \left(\frac{1}{2N}\sum_{i=1}^{N} \left(\hat{\varepsilon}_i X_i - \hat{\varepsilon}_{\ell_{X_1}(i)} X_{\ell_{X_1}(i)}\right)\left(\hat{\varepsilon}_i X_i - \hat{\varepsilon}_{\ell_{X_1}(i)} X_{\ell_{X_1}(i)}\right)'\right) \cdot \left(\frac{1}{N}\sum_{i=1}^{N} X_i X_i'\right)^{-1}.$$

This estimator is consistent for the conditional variance $\mathbb{V}_{\text{cond}}$. In contrast, replacing $\hat{\varepsilon}_{\ell_X(i)}$ by $\hat{\varepsilon}_{\ell_{X_1}(i)}$ in the expression for $\widetilde{\mathbb{V}}_{\text{cond}}$ would not lead to a consistent estimator for the variance. Although the asymptotic variance $\mathbb{V}_{\text{cond}}$ is less than or equal to the EHW variance $\mathbb{V}_{\text{pop}}$, this need not hold for the estimators. In finite samples it may well be the case that $\widehat{\mathbb{V}}_{\text{cond}}$ is larger than $\widehat{\mathbb{V}}_{\text{pop}}$. We study the finite sample behavior of the variance estimator in a simulation study in Section 3.6.

In the remainder of this paper we will generalize the results in this section to maximum likelihood and method of moments settings, and state formal results concerning the large sample properties of the variance estimators. In the general settings the estimators are no longer least squares estimators, and we will modify the terminology to reflect this. We will use θ_{pop} for population estimands that generalize the best linear predictor θ_{pop} in the regression case, and θ_{cond} for the conditional version that generalizes the conditional best linear predictor θ_{cond} in the regression case.

3.3 Motivation for Conditional Estimands

In this section we address the question whether, when, and why the estimand conditional on the covariates may be of interest. We emphatically do not wish to argue that the conditional estimand is the appropriate object of interest in all cases. Rather, we wish to make the case, through two examples, that it depends on the context what the appropriate object is, and that in some settings the conditional best linear predictor is more appropriate than the standard unconditional estimand.

One way to frame the question is in terms of different repeated sampling perspectives one can take. We can consider the distribution of the least squares estimator over repeated samples where we redraw the pairs X_i and Y_i (the random regressor case), or we can consider the distribution over repeated samples where we keep the values of X_i fixed and only redraw the Y_i (the fixed regressor case). Under general misspecification both the mean and variance of these two distributions of the estimator will differ. The population estimand θ_{pop} is the approximate (in a large sample sense) average over the repeated samples when we redraw both X_i and Y_i, and $\theta_{\text{cond}}(\mathbf{X})$ is the approximate average over the repeated samples where X_i is held fixed. Many introductory treatments of regression analysis briefly introduce the fixed and random regressor concepts, with a variety of opinions on what the most relevant perspective is. Wooldridge writes that "reliance on fixed regressors ... can have unintended consequences. ... Because our focus is on asymptotic analysis, we have the luxury of allowing for random explanatory variables throughout the book" (Wooldridge, 2002, p.10-11). Goldberger (1991) takes a different position, assuming "$\mathbf{X}$ nonstochastic, which says that the elements of $\mathbf{X}$ are constants, that is, degenerate random variables. Their values are fixed in repeated samples ..." (Goldberger, p. 164). Van der Vaart (2000) writes "We assume that the independent variables are a random sample in order to fit the example in our i.i.d. notation, but the analysis could be carried out conditionally as well." (VanderVaart, p. 57), and Gelman and Hill (2006) focus on the fixed regressor perspective, writing "This book follows the usual approach of setting up regression models in the measurement-error framework ($y = a + bx + \epsilon$), with the sampling interpretation implicit

in that the errors $\epsilon_1, \ldots, \epsilon_n$, can be considered as a random sample from a distribution" (Gelman and Hill, p.17). These discussions are in the context of correctly specified regression models, however, where the averages of the distributions under the two repeated sampling perspectives coincide, and their variances agree in large samples. A point that has not received attention in the literature is that under general misspecifiaction, the random versus fixed regressor distinction has implications for inference that do not vanish with the sample size.

Another point is that the sole difference between the population and conditional estimands is the weight function used to measure the difference between the model and the true data generating process. For the population estimand the weight function depends on the population distribution of the potential conditioning variables, and for the conditional estimand it is the sample distribution of these variables. Because the population distribution of these variables, unlike the sample distribution, is unknown, in general there is more uncertainty about the population estimand. Thus, focusing on the conditional estimand θ_{cond} generally leads to smaller standard errors than focusing on the population estimand θ_{pop}.

Example I (Convenience Sample)

In the first example we want to make the case that sometimes there is intrinsically no more interest in θ_{pop} than θ_{cond} because neither the weighting scheme corresponding to the population distribution, nor the weighting scheme corresponding to the empirical distribution function, is obviously of primary interest.

Consider the study of lottery winners by Imbens *et al.* (2001). Imbens, Rubin and Sacerdote surveyed individuals who won large prizes in the lottery. Using a standard life-cycle model of labor supply they focus on linear regressions of subsequent labor earnings on the annual prize and some additional covariates including prior earnings. The coefficient on the prize in this linear regression can be interpreted as the marginal propensity to consume out of unearned income, an economically meaningful parameter (e.g., Pencavel, 1986). Even if the conditional expectation as a function of the prize is nonlinear, it may still

be interesting to focus on the coefficient in the linear regression, partly because it facilitates comparison across studies. The question is whether the linear approximation should be based on weighting the squared difference between the true regression function and the linear predictor by the population or empirical distribution of lottery prizes. There does not appear to be a strong substantive argument for preferring one weighting function (and thus the corresponding estimand) over the other. □

Example II (Experimental Design)

Karlan and List (2007) carried out an experimental evaluation of incentives for charitable giving. Among the results Karlan and List report are probit regression estimates where the object of interest is the regression coefficient on the indicator for being offered a matching incentive for charitable giving. The specification of the probit regression function also includes characteristics of the matching incentives.

In this case the difference between $\mathbb{V}_{\text{pop}}$ and $\mathbb{V}_{\text{cond}}$ is that $\mathbb{V}_{\text{pop}}$ takes into account sampling variation in $\hat{\theta}$ due to variation in the sample values of the matching incentives over the repeated samples, whereas $\mathbb{V}_{\text{cond}}$ conditions on these values. Given that the distribution of these incentives in this experiment is fixed by the researchers there appears to be no reason to take this uncertainty into account, and we submit that the appropriate measure of uncertainty is $\mathbb{V}_{\text{cond}}$ rather than $\mathbb{V}_{\text{pop}}$. □

3.4 Inference for Conditional Estimands

In this section we present the main formal results of the paper, covering linear regression, maximum likelihood, and method of moments estimators. We cover settings where we condition on the full set of regressors as well as cases where we condition on a subset of the regressors. We focus on the just-identified case, although the results can be extended to over-identified generalized method of moments settings, for example using empirical likelihood approaches (e.g., Qin and Lawless, 1994; Imbens, 1997; Imbens *et al.*, 1998; and Newey and Smith, 2004).

Suppose we have a random sample of size N of a pair of random vectors, (X_i, Y_i),

$i = 1, \ldots, N$. Let $\mathbf{X}$ and $\mathbf{Y}$ be the $N \times K_X$ and $N \times K_Y$ matrices with i-th rows equal to X_i' and Y_i' respectively. The distinction between X_i and Y_i is that we may wish to condition on the X_i in defining the estimand. We are interested in a finite dimensional parameter θ, defined in general as some function of the joint distribution of (X_i, Y_i). Under some statistical model it follows that

$$\mathbb{E}\left[\psi(Y_i, X_i, \theta)\right] = 0, \tag{3.4.1}$$

with the dimension of θ equal to that of ψ. The model may have additional implications beyond this moment restriction, but these are not used for estimation. For example, it may be the case that the conditional moment has expectation zero,

$$\mathbb{E}\left[\,\psi(Y_i, X_i, \theta)|\, X_i\right] = 0.$$

Alternatively, we may have specified the joint distribution of Y_i and X_i, in which case $\psi(y, x, \theta)$ could equal to the score function. In that case the model has the additional implication that minus the expected value of the derivatives of $\psi(y, x, \theta)$ with respect to θ is equal to the expected value of the second moments of $\psi(y, x, \theta)$. Based only on (3.4.1), and not on any other implications of the motivating model, we may wish to estimate θ by $\hat{\theta}$, which satisfies

$$\frac{1}{N}\sum_{i=1}^{N} \psi(Y_i, X_i, \hat{\theta}) = 0.$$

We are interested in the properties of the estimator $\hat{\theta}$ under general misspecification of the model that motivated the moment restriction.

The standard approach to generalized method of moments (GMM) and empirical likelihood estimation (Hansen, 1984; Newey and McFadden, 1994; Wooldridge, 2002; Qin and Lawless, 1993; Imbens, Johnson and Spady, 1997) focuses on the value θ_{pop} that solves

$$\mathbb{E}\left[\psi(Y_i, X_i, \theta_{\text{pop}})\right] = 0.$$

If the pairs (X_i, Y_i), for $i = 1, \ldots, N$ are independent and identically distributed, then under

regularity conditions,

$$\sqrt{N}\left(\hat{\theta}-\theta_{\text{pop}}\right) \xrightarrow{d} \mathcal{N}\left(0, \mathbb{V}_{\text{gmm,pop}}\right), \qquad \text{where } \mathbb{V}_{\text{gmm,pop}} = \left(\Gamma'\Delta_{\text{pop}}^{-1}\Gamma\right)^{-1},$$

with

$$\Gamma = \mathbb{E}\left[\frac{\partial}{\partial\theta'}\psi(Y_i, X_i, \theta_{\text{pop}})\right], \qquad \text{and } \Delta_{\text{pop}} = \mathbb{E}\left[\psi(Y_i, X_i, \theta_{\text{pop}})\psi(Y_i, X_i, \theta_{\text{pop}})'\right].$$

Now we focus on the conditional estimand, where we condition on $\mathbf{X}$. Define $\theta_{\text{cond}}(\mathbf{X})$ as the solution to

$$\mathbb{E}\left[\sum_{i=1}^{N}\psi(Y_i, X_i, \theta)\middle|\mathbf{X}\right] = 0. \tag{3.4.2}$$

Note that implicitly θ_{cond} is a function of $\mathbf{X}$. If the original model implied that the conditional expectation of $\psi(Y_i, X_i, \theta)$ given X_i is equal to zero, then $\theta_{\text{cond}} = \theta_{\text{pop}}$, but this need not hold in general. The motivation for the estimand is the same as in the best-linear-predictor case. In cases where the model implies a conditional moment restriction, but we are concerned about misspecification, we may wish to focus on the value for θ that minimizes the discrepancy between $\mathbb{E}[\psi(Y_i, X_i, \theta)|X_i]$ and zero. We can weight the discrepancy by the population distribution of the X_i's, or by the empirical distribution. The conditional estimand corresponds to the case where the weights are based on the empirical distribution function.

We make the following assumptions. These are closely related to standard assumptions used for establishing asymptotic properties for moment-based estimators. See for example Newey and McFadden (1994).

Assumption 3.1 *(Y_i, X_i), for $i = 1, \ldots, N$, are independent and identically distributed.*

Assumption 3.2 *(i) For some compact $\Theta \subset \mathbb{R}^K$, there is a unique value, $\theta_{\text{pop}} \in \Theta$, such that $\mathbb{E}[\psi(Y_i, X_i, \theta_{\text{pop}})] = 0$; (ii) $\psi(Y, X, \theta)$ is continuous at each $\theta \in \Theta$ with probability one; (iii) $\mathbb{E}[\sup_{\theta\in\Theta}\|\psi(Y_i, X_i, \theta)\|] < \infty$.*

Theorem 3.1 *If Assumptions 3.1 and 3.2 hold, then:*

$$\hat{\theta} - \theta_{\text{pop}} \xrightarrow{p} 0,$$

and

$$\hat{\theta} - \theta_{\text{cond}}(\mathbf{X}) \xrightarrow{p} 0.$$

All proofs are given in the appendix.

Assumption 3.3 *(i)* θ_{pop} *is an interior point of* Θ*; (ii)* $\psi(y, x, \theta)$ *is continuously differentiable with respect to* θ *in an open neighborhood* $\mathcal{B}$ *of* θ_{pop}*; (iii)* $\mathbb{E}[\|\psi(Y_i, X_i, \theta_{\text{pop}})\|^2] < \infty$*; (iv)* $\mathbb{E}[\sup_{\theta \in \mathcal{B}} \|\partial\psi(Y_i, X_i, \theta)/\partial\theta'\|] < \infty$*; (v)* $\Gamma = \mathbb{E}[\partial\psi(Y_i, X_i, \theta_{\text{pop}})/\partial\theta']$ *is nonsingular.*

Theorem 3.2 *Under Assumptions 3.1-3.3,*

$$\sqrt{N}(\hat{\theta} - \theta_{\text{pop}}) \xrightarrow{d} N(0, \Gamma^{-1}\Delta_{\text{pop}}(\Gamma^{-1})'),$$

where $\Delta_{\text{pop}} = \mathbb{E}[\psi(Y_i, X_i, \theta_{\text{pop}})\psi(Y_i, X_i, \theta_{\text{pop}})']$ *and*

$$\sqrt{N}(\hat{\theta} - \theta_{\text{cond}}(\mathbf{X})) \xrightarrow{d} N(0, \Gamma^{-1}\Delta_{\text{cond}}(\Gamma^{-1})'),$$

where $\Delta_{\text{cond}} = \mathbb{E}[\mathbb{V}(\psi(Y_i, X_i, \theta_{\text{pop}})|X_i)]$.

Corollary 3.1 *Under the conditions of Theorem 3.2, if* $\mathbb{E}[\psi(Y_i, X_i, \theta_{\text{pop}})|X_i = x] = 0$ *for almost all* x *in the support of* X_i*, then* $\sqrt{N}(\hat{\theta} - \theta_{\text{pop}})$ *and* $\sqrt{N}(\hat{\theta} - \theta_{\text{cond}}(\mathbf{X}))$ *have the same asymptotic distribution.*

Assumption 3.3*(ii)* requires differentiability of $\psi(y, x, \theta)$. This assumption can, however, be replaced by asymptotic equicontinuity conditions as in Huber (1967), Pakes and Pollard (1989), Andrews (1994), or Newey and McFadden (1994). In Appendix C.2 we show that the results of Theorem 3.2 and Corollary 3.1 hold under an asymptotic equicontinuity condition, with the only change that for the non-differentiable case we have $\Gamma = \partial\mathbb{E}[\psi(Y_i, X_i, \theta_{\text{pop}})]/\partial\theta'$. Example VI below discusses the case of L_1 (quantile) regression. Notice that the consistency result in Theorem 3.1 does not require everywhere differentiability of $\psi(y, x, \theta)$.

We now discuss two additional examples that illustrate the differences between the large sample variances of $\sqrt{N}(\hat{\theta} - \theta_{\text{pop}})$ and $\sqrt{N}(\hat{\theta} - \theta_{\text{cond}}(\mathbf{X}))$. The first example is related to the discussion in Chow (1984).

Example III (Maximum Likelihood Estimation)

Suppose we specify the conditional distribution of Y_i given X_i as $f(y|x;\theta)$. We estimate the model by maximum likelihood:

$$\hat{\theta} = \arg\max_{\theta} \sum_{i=1}^{N} \ln f(Y_i|X_i;\theta).$$

The normalized asymptotic variance under correct specification, and under some regularity conditions, is equal to the inverse of the information matrix $\mathcal{I}_\theta^{-1}$, where

$$\mathcal{I}_\theta = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta\partial\theta'} \ln f(Y_i|X_i;\theta)\right] = \mathbb{E}\left[\frac{\partial}{\partial\theta} \ln f(Y_i|X_i;\theta) \cdot \frac{\partial}{\partial\theta} \ln f(Y_i|X_i;\theta)'\right].$$

Huber (1967) and White (1982) analyzed the properties of the maximum likelihood estimator under general misspecification of the conditional density. Let

$$\theta_{\text{pop}} = \arg\max_{\theta} \mathbb{E}\left[\ln f(Y_i|X_i;\theta)\right].$$

They showed that under general misspecification,

$$\hat{\theta} \xrightarrow{p} \theta_{\text{pop}}, \qquad \text{and} \quad \sqrt{N}\cdot\left(\hat{\theta} - \theta_{\text{pop}}\right) \xrightarrow{d} \mathcal{N}\left(0, \Gamma^{-1}\Delta_{\text{pop}}\Gamma^{-1}\right),$$

with

$$\Gamma = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta\partial\theta'} \ln f(Y_i|X_i;\theta_{\text{pop}})\right], \quad \Delta_{\text{pop}} = \mathbb{E}\left[\frac{\partial}{\partial\theta} \ln f(Y_i|X_i;\theta_{\text{pop}}) \cdot \frac{\partial}{\partial\theta} \ln f(Y_i|X_i;\theta_{\text{pop}})'\right].$$

The conditional version of the estimand under general misspecification is

$$\theta_{\text{cond}}(\mathbf{X}) = \arg\max_{\theta} \sum_{i=1}^{N} \mathbb{E}\left[\ln f(Y_i|X_i;\theta)\middle|\, X_i\right],$$

where the expectation is taken only over the conditional distribution of Y_i given X_i. Theorem 3.2 implies that

$$\sqrt{N}\cdot\left(\hat{\theta} - \theta_{\text{cond}}(\mathbf{X})\right) \xrightarrow{d} \mathcal{N}\left(0, \Gamma^{-1}\Delta_{\text{cond}}\Gamma^{-1}\right),$$

where

$$\Delta_{\text{cond}} = \mathbb{E}\left[\mathbb{V}\left(\frac{\partial}{\partial\theta} \ln f(Y_i|X_i,\theta_{\text{pop}})\middle|\, X_i\right)\right].$$

If the model is correctly specified, then $\Delta_{\text{pop}} = \Delta_{\text{cond}}$. If the model is misspecified, then

$$\mathbb{E}\left[\frac{\partial}{\partial\theta}\ln f(Y_i|X_i,\theta_{\text{pop}})\right] = 0, \quad \mathbb{E}\left[\frac{\partial}{\partial\theta}\ln f(Y_i|X_i,\theta_{\text{pop}})\middle| X_i = x\right] \neq 0,$$

for x in a set of positive probability. For such x,

$$\mathbb{E}\left[\frac{\partial}{\partial\theta}\ln f(Y_i|X_i;\theta_{\text{pop}})\cdot\frac{\partial}{\partial\theta}\ln f(Y_i|X_i;\theta_{\text{pop}})'\middle| X_i = x\right] \geq \mathbb{V}\left(\frac{\partial}{\partial\theta}\ln f(Y_i|X_i,\theta_{\text{pop}})\middle| X_i = x\right),$$

implying that in general $\Delta_{\text{pop}} - \Delta_{\text{cond}}$ is positive semi-definite. □

Example IV (Quantile Regression)

Suppose that the τ-th conditional quantile of Y_i given X_i is a linear function, so $\mathbb{E}[I_{[Y_i \leq X_i'\theta_{\text{pop}}]}|X_i = x] = \tau$, where I_A is the indicator function for the event A. Therefore, $\mathbb{E}[X_i(I_{[Y_i \leq X_i'\theta_{\text{pop}}]} - \tau)] = 0$. The quantile regression estimator $\hat{\theta}$ (Koenker and Bassett Jr, 1978) solves the analogous sample moment restrictions:

$$\left\| \frac{1}{N}\sum_{i=1}^{N} X_i(I_{[Y_i \leq X_i'\hat{\theta}]} - \tau) \right\| = o_p(1/\sqrt{N}) \tag{3.4.3}$$

(see Powell, 1984). If the quantile regression model is misspecified, so $\mathbb{E}[I_{[Y_i \leq X_i'\theta_{\text{pop}}]}|X_i = x] \neq \tau$ for some x in a set of positive probability, there will generally still be a value θ_{pop} that solves (3.4.3). Under regularity conditions the quantile regression estimator estimates that parameter, and its distribution is

$$\sqrt{N}(\hat{\theta} - \theta_{\text{pop}}) \xrightarrow{p} \mathcal{N}(0, \Gamma^{-1}\Delta_{\text{pop}}\Gamma^{-1}),$$

where

$$\Gamma = \mathbb{E}[f_{Y|X=X_i}(X_i'\theta_{\text{pop}})X_iX_i']$$

and

$$\Delta_{\text{pop}} = \mathbb{E}[X_i(I_{[Y_i - X_i'\theta_{\text{pop}} \leq 0]} - \tau)^2X_i']$$

(see, for example, Angrist *et al.* (2006), or Appendix C.3). Angrist, Chernozhukov, and Fernández-Val (2006) provide an interpretation of quantile regression under misspecification.

In Appendix C.3 we show that, in addition:

$$\sqrt{N}(\hat{\theta} - \theta(\mathbf{X})) \xrightarrow{p} \mathcal{N}(0, \Gamma^{-1}\Delta_{\text{cond}}\Gamma^{-1}),$$

where

$$\Delta_{\text{cond}} = \mathbb{E}[X_i \mathbb{V}(I_{[Y_i - X_i'\theta \leq 0]}|X_i)X_i'].$$

Because $\mathbb{E}[(I_{[Y_i - X_i'\theta_{\text{pop}} \leq 0]} - \tau)^2|X_i] \geq V(I_{[Y_i - X_i'\theta \leq 0]}|X_i)$, it follows that $\Delta_{\text{pop}} - \Delta_{\text{cond}}$ is positive semi-definite. Under correct specification, $\mathbb{E}[I_{[Y_i - X_i'\theta \leq 0]}|X_i] = \tau$, so $\mathbb{E}[(I_{[Y_i - X_i'\theta \leq 0]} - \tau)^2|X_i] = \mathbb{V}(I_{[Y_i - X_i'\theta \leq 0]}|X_i) = \tau(1-\tau)$ and $\Delta_{\text{cond}} = \Delta_{\text{pop}}$. □

Next, we consider estimation of the variance in the general case. Estimation of Γ is the same as for the population estimand:

$$\hat{\Gamma} = \frac{1}{N}\sum_{i=1}^{N} \frac{\partial}{\partial \theta'}\psi(Y_i, X_i, \hat{\theta}).$$

The key question concerns estimation of Δ_{cond}. Our proposed estimator matches each unit to the closest unit in terms of X_i, and then differences the values of the moment function:

$$\hat{\Delta}_{\text{cond}} = \frac{1}{2N}\sum_{i=1}^{N} \left(\psi(Y_i, X_i, \hat{\theta}) - \psi(Y_{\ell_X(i)}, X_{\ell_X(i)}, \hat{\theta})\right)\left(\psi(Y_i, X_i, \hat{\theta}) - \psi(Y_{\ell_X(i)}, X_{\ell_X(i)}, \hat{\theta})\right)',$$

where $\ell_X(i)$ is as defined in (3.2.10). We then combine these estimates to get an estimator for the variance of the conditional estimand:

$$\widehat{\mathbb{V}}_{\text{gmm,cond}} = \hat{\Gamma}^{-1}\hat{\Delta}(\hat{\Gamma}^{-1})'.$$

Assumption 3.4 *The support of X_i is compact. The conditional expectation $\mathbb{E}[\psi^k(Y_i, X_i, \theta)|X_i = x]$ is Lipschitz in x with constant C_k for $k \leq 4$, for all θ in an open neighborhood of θ_{pop}, where C_k does not depend on θ.*

Theorem 3.3 (Conditional Variance for Method of Moments Estimators) *Suppose Assumptions 3.1-3.4 hold. Then*

$$\widehat{\mathbb{V}}_{\text{gmm,cond}} \xrightarrow{p} \mathbb{V}_{\text{gmm,cond}}.$$

3.5 An Application to Cross-Country Growth Regressions

For an illustration of the methods discussed in this paper we turn to an analyis in Sachs and Warner (1997) of the determinants of country-level growth rates. Sachs and Warner have data for 83 countries on the county's *per capita* growth rate between 1965 and 1990, and wish to relate this outcome to country-level fiscal policies. These policies include the degree of openness of the country ("open") and the central government budget balance ("cgb"). Sachs and Warner estimate a linear regression of the *per capita* growth rate on these variables, also including a number of characteristics of the country such as its location relative to the tropics and the sea (landlocked or not), and some measures of the economic conditions at the beginning of this period, including gross domestic product in 1965 ("gdp65").

The estimates are reported in Table 3.1, with the variables described in Table 3.2. We calculate the EHW standard errors, as well as our proposed conditional standard errors where the variables we condition on include all characteristics of the countries other than the economic policy variables open, open$\times$gdp65, and cgb which are directly under the control of the government. It would appear reasonable that at least some of these variables should be conditioned on, including whether a country is landlocked and what share of its landmass is in the tropics.

We find that the standard errors for the key variables, the indicator for being open and its interaction with gdp in 1965 go down by about 7%.

3.6 Two Simulation Studies

In this section we assess the small sample properties of the variance estimators. We focus on two models, first a linear regression and second a logistic regression model.

3.6.1 A Simulation Study of a Linear Model

We consider estimating a regression function with K regressors. the first regressor, X_{1i}, has a mixture of a normal distribution with mean zero and unit variance, and a log normal

Table 3.1: *Cross Country Growth Regression, Dependent variable: per capita GDP growth between 1965 and 1990*

	$\hat{\beta}$	$\sqrt{\hat{\mathbb{V}}_{\text{pop}}}$	$\sqrt{\hat{\mathbb{V}}_{\text{cond}}}$
constant	1.66	3.08	3.03
gdp65	-1.50	0.18	0.17
open	10.91	2.76	2.56
open65	-1.08	0.35	0.33
dpop	0.69	0.40	0.45
cgb	0.115	0.025	0.023
inst	0.315	0.071	0.068
tropics	-0.83	0.25	0.24
land	-0.58	0.21	0.26
sxp	-3.92	1.22	1.21
life	0.35	0.12	0.12
life2	-0.003	0.001	0.001

$N = 83$

$R^2 = 0.862$

Table 3.2: *Description of Variables: Cross Country Growth Regression*

Variable	Description
dependent variable	Average annual growth in real GDP per economically active population between 1970 and 1989.
gdp65	Log of real GDP per economically active population in 1965.
open	Fraction of years during the period 1965-1990 in which the country is rated as an open economy according to the criteria in Sachs and Warner (1995).
open65	open*gdp65.
dpop	Difference between the growth rate of the economically active population (between ages 15 and 65) and growth of total population.
cgb	Current revenues minus current expenditures of the central government, expressed as a fraction of GDP.
inst	Institutional quality index.
tropics	Approximate proportion of land area subject to a tropical climate.
land	Dummy variable that equals one if a country is landlocked.
sxp	Share of exports of primary products in GNP in 1970.
life	Life expectancy at birth, circa 1965-1970.
life2	Life squared.

distribution with parameters $\mu = 0$ and $\sigma^2 = 0.5$. The mixture probability for the log normal component is p. We use two values for p in the simulations, $p = 0$ and $p = 0.1$ with the latter corresponding to a design with high leverage covariates. The remaining $K - 1$ covariates have normal distributions with mean zero and unit variance. All covariates are independent. We use two values for the number of covariates: $K = 1$ where only X_{1i} is present in the regression function, and $K = 5$ where there are four additional regressors. We use two sample sizes, $N = 50$ and $N = 200$. The conditional distribution of Y_i given $(X_{1i}, \ldots, X_{Ki})$ is Normal:

$$Y_i | X_{1i}, \ldots, X_{Ki} \sim \mathcal{N}\left(\mu_i, \sigma_i^2\right),$$

where

$$\mu_i = X_{1i} + \delta \cdot \left(X_{1i}^2 - 1\right), \text{and} \quad \ln \sigma_i^2 = 1 - \gamma \cdot X_{1i}.$$

A non-zero value for δ makes the model nonlinear and implies that the linear regression model is misspecified. We use two values for δ. In the first design we fix $\delta = 0$ (correct specification), and in the second design we use a larger value, $\delta = 1$ (misspecification). A non-zero value for γ implies heteroskedasticity. We use two values for γ, $\gamma = 0$ (homoskedasticity) and $\gamma = 0.5$ (heteroskedasticity). With two values for each of five parameters of the design, $p \in \{0, 0.1\}$, $K \in \{1, 5\}$, $N \in \{50, 200\}$, $\delta \in \{0, 0.1\}$, and $\gamma \in \{0.0.5\}$, we consider a total of 32 designs.

For each of the 32 designs we focus on estimating a linear regression function

$$Y_i = \theta_0 + \sum_{k=1}^{K} \theta_k \cdot X_{ki} + \varepsilon_i.$$

Table 3.3 presents the results, based on 50,000 replications for each design. We focus on the coefficient on X_{1i}, denoted by θ (dropping the subscript 1 for ease of notation). For all designs we report four coverage rates. First the coverage frequency of the conventional (EHW standard error based) 95% confidence interval for θ_{pop}. This coverage frequency is calculated as the frequency with which $(\hat{\theta} - \theta_{\text{pop}}) / \sqrt{\hat{\mathbb{V}}_{\text{pop}}}$ is less than 1.96 in absolute value. Note that both θ_{pop} and θ_{cond} need to be numerically evaluated for these data generating processes. The nominal coverage rate of the confidence intervals is 0.95. Next, the frequency

with which the same confidence interval covers θ_{cond}, that is, the frequency with which $(\hat{\theta} - \theta_{\text{cond}}(\mathbf{X}))/\sqrt{\hat{\mathbb{V}}_{\text{pop}}}$ is less than 1.96 in absolute value. This should in large samples be at least 0.95, and more than 0.95 in misspecified models according to our formal results. We also report the coverage rates for confidence intervals based on the conditional standard errors. Now the coverage for θ_{pop} could be less than 0.95, but the coverage for $\theta_{\text{cond}}(\mathbf{X})$ should be 0.95.

In the first design (Design I) with a single covariate, 50 observations, a linear conditional expectation and a normal regressor and homoskedasticity, both variance estimators lead to coverage rates around 92-93%, with the EHW variance doing slightly better. With five covariates (Design II), the difference between the two variance estimators (in favor of the EHW variance estimator) becomes more pronounced. Having a skewed distribution for the covariate with some high leverage values does not change the coverage rates very much in Design III. With 200 observations (Design V), the coverage rates become closer to the nominal coverage rates for both variance estimators. Given heteroskedasticity (Design IX), the EHW variance estimator does substantially better with a coverage rate of 91%, whereas the conditional variance estimator leads to confidence intervals with a coverage rate of 88% Allowing for misspecification of the regression function (Design XVII) changes the coverage rates substantially. The coverage rate, based on the EHW estimator, for θ_{pop}, is 90%. The coverage rate based on the conditional variance estimator, for θ_{cond}, is much closer to the nominal level, at 0.94.

Over the 32 designs, the worst performance of the EHW variance estimator is in Design XX, with misspecification and high leverage covariates, 50 observations, and 5 covariates, where the coverage rate is 79% instead of 95%. The worst performance of the conditional variance estimator is in Design XII, with a linear model, heteroskedasticity, five covariates, with high leverage, and 50 observations, with an actual coverage rate of 88%. It appears that the conditional variance estimator is more sensitive to heteroskedasticity, but less sensitive to the distribution of the covariates. Overall the worst case for the conditional variance estimator is substantially better than for the EHW variance estimator.

Table 3.3: *Coverage Rate 95% Confidence Interval and Median Estimated Standard Error (Linear Model, 50,000 Replications)*

Estimand ⟶						θ_{pop}		θ_{cond}		median	
Variance ⟶						$\hat{\mathbb{V}}_{\text{pop}}$	$\hat{\mathbb{V}}_{\text{cond}}$	$\hat{\mathbb{V}}_{\text{pop}}$	$\hat{\mathbb{V}}_{\text{cond}}$	$\sqrt{\hat{\mathbb{V}}_{\text{pop}}}$	$\sqrt{\hat{\mathbb{V}}_{\text{cond}}}$
	Misspec	Homo	Samp Size	High Lev	K						
I	No	Yes	50	No	1	0.926	0.921	0.926	0.921	0.367	0.368
II	No	Yes	50	No	5	0.912	0.897	0.912	0.897	0.366	0.354
III	No	Yes	50	Yes	1	0.923	0.916	0.923	0.916	0.345	0.344
IV	No	Yes	50	Yes	5	0.907	0.892	0.907	0.892	0.344	0.331
V	No	Yes	200	No	1	0.945	0.941	0.945	0.941	0.190	0.188
VI	No	Yes	200	No	5	0.940	0.927	0.940	0.927	0.190	0.182
VII	No	Yes	200	Yes	1	0.942	0.936	0.942	0.936	0.177	0.175
VIII	No	Yes	200	Yes	5	0.939	0.925	0.939	0.925	0.177	0.169
IX	No	No	50	No	1	0.914	0.877	0.914	0.877	0.561	0.548
X	No	No	50	No	5	0.893	0.853	0.893	0.853	0.542	0.508
XI	No	No	50	Yes	1	0.921	0.879	0.921	0.879	0.508	0.493
XII	No	No	50	Yes	5	0.901	0.861	0.901	0.861	0.492	0.460

Continued on next page

Table 3.3: *(continued)*

Estimand ⟶						θ_{pop}		θ_{cond}		median	
Variance ⟶						$\hat{\mathbb{V}}_{\text{pop}}$	$\hat{\mathbb{V}}_{\text{cond}}$	$\hat{\mathbb{V}}_{\text{pop}}$	$\hat{\mathbb{V}}_{\text{cond}}$	$\sqrt{\hat{\mathbb{V}}_{\text{pop}}}$	$\sqrt{\hat{\mathbb{V}}_{\text{cond}}}$
	Misspec	Homo	Samp Size	High Lev	K						
XIII	No	No	200	No	1	0.938	0.915	0.938	0.915	0.318	0.310
XIV	No	No	200	No	5	0.937	0.903	0.937	0.903	0.316	0.291
XV	No	No	200	Yes	1	0.943	0.917	0.943	0.917	0.284	0.276
XVI	No	No	200	Yes	5	0.940	0.904	0.940	0.904	0.282	0.260
XVII	Yes	Yes	50	No	1	0.904	0.811	0.978	0.938	0.503	0.397
XVIII	Yes	Yes	50	No	5	0.885	0.826	0.967	0.941	0.489	0.422
XIX	Yes	Yes	50	Yes	1	0.816	0.673	0.984	0.948	0.535	0.404
XX	Yes	Yes	50	Yes	5	0.789	0.695	0.976	0.954	0.516	0.434
XXI	Yes	Yes	200	No	1	0.938	0.806	0.993	0.948	0.278	0.195
XXII	Yes	Yes	200	No	5	0.934	0.845	0.991	0.964	0.276	0.215
XXIII	Yes	Yes	200	Yes	1	0.796	0.569	0.998	0.965	0.333	0.204
XXIV	Yes	Yes	200	Yes	5	0.791	0.627	0.997	0.980	0.329	0.233

Continued on next page

Table 3.3: *(continued)*

Estimand ⟶						θ_{pop}		θ_{cond}		median	
Variance ⟶						$\hat{\mathbb{V}}_{\text{pop}}$	$\hat{\mathbb{V}}_{\text{cond}}$	$\hat{\mathbb{V}}_{\text{pop}}$	$\hat{\mathbb{V}}_{\text{cond}}$	$\sqrt{\hat{\mathbb{V}}_{\text{pop}}}$	$\sqrt{\hat{\mathbb{V}}_{\text{cond}}}$
	Misspec	Homo	Samp Size	High Lev	K						
XXV	Yes	No	50	No	1	0.892	0.827	0.937	0.887	0.655	0.567
XXVI	Yes	No	50	No	5	0.870	0.819	0.922	0.884	0.628	0.556
XXVII	Yes	No	50	Yes	1	0.871	0.763	0.950	0.903	0.675	0.561
XXVIII	Yes	No	50	Yes	5	0.841	0.757	0.939	0.904	0.644	0.555
XXIX	Yes	No	200	No	1	0.931	0.865	0.966	0.919	0.376	0.314
XXX	Yes	No	200	No	5	0.928	0.868	0.964	0.924	0.373	0.312
XXXI	Yes	No	200	Yes	1	0.878	0.727	0.982	0.941	0.423	0.318
XXXII	Yes	No	200	Yes	5	0.873	0.739	0.981	0.950	0.418	0.324

3.6.2 A Simulation Study of a Logistic Regression Model

Next, we do a similar simulation study in a nonlinear setting. Here the outcome is a binary indicator. We estimate a logistic regression model specified as:

$$\text{pr}(Y_i = 1|X_{1i}, \ldots, X_{Ki}) = \frac{1}{1 + \exp(\theta_0 + \sum_{k=1}^{K} \theta_k \cdot X_{ki})}.$$

The data are generated through a model where a latent index Y_i^* satisfies

$$Y_i^* = \theta_0 + \sum_{k=1}^{K} \theta_k \cdot X_{ki} + \varepsilon_i,$$

and the observed outcome is the indicator that Y_i^* is non-negative:

$$Y_i = I_{[Y_i^* \geq 0]}.$$

In the base case, there are 50 observations, and ε_i has a logistic distribution so that the logistic regression model is correctly specified. In this case there is a single covariate ($K = 1$), $\theta_1 = 1$, $\theta_0 = 0$, and the covariate has a standard Normal distribution with unit variance.

We can consider combinations of five modifications, similar to those in the linear model. First, we allow for the presence of four additional covariates ($K = 5$), with the additional covariates all having independent normal distributions with zero coefficients. Second, we change the distribution of the first covariate to include high leverage points by making it a mixture of a standard Normal distribution and a log normal distribution with parameters 0 and 0.5, and the probability of the log normal component equal to 0.1. Third, we change the sample size to 200. Fourth, we multiply the ε_i for all units by $\exp(1 - 0.5 \cdot X_{1i})$. In the linear case this corresponds to introducing heteroskedasticity, but here this also implies misspecification of the logistic regression model. Finally, we directly misspecify the regression function by changing the specification of Y_i^* to

$$Y_i^* = X_{1i} + \left(X_{1i}^2 - 1\right) + \varepsilon_i.$$

Table 3.4 presents the results for the 32 designs generated as combinations of these changes to the base design, based on 50,000 replications. There are some qualitative

differences with the simulations for the linear case. There are generally bigger differences between the two variance estimators, $\hat{\mathbb{V}}_{\text{cond}}$ and $\hat{\mathbb{V}}_{\text{pop}}$. The coverage rates for confidence intervals, for θ_{pop} based on $\hat{\mathbb{V}}_{\text{pop}}$, and for $\theta_{\text{cond}}(\mathbf{X})$ based on $\hat{\mathbb{V}}_{\text{cond}}$, are closer to nominal levels. In contrast, inference for for $\theta_{\text{cond}}(\mathbf{X})$ based on $\hat{\mathbb{V}}_{\text{pop}}$ leads to confidence intervals with substantially higher coverage, and inference for θ_{cond} based on $\hat{\mathbb{V}}_{\text{cond}}(\mathbf{X})$ leads to substantial undercoverage.

In general, inference for $\theta_{\text{cond}}(\mathbf{X})$ is less affected by the changes in the design than inference for θ_{pop}. For example, the worst design for θ_{pop} is still Design XX, with both misspecification and high leverage covariates, where the coverage rate is 0.930. For the conditional estimand, the worst designs are those with misspecification, with coverage rates around 0.924, still close to the nominal 0.95 level.

Table 3.4: *Coverage Rate 95% Confidence Interval and Median Estimated Standard Error (Logistic Model, 50,000 Replications)*

Estimand ⟶						θ_{pop}		θ_{cond}		median	
Variance ⟶						$\hat{\mathbb{V}}_{\text{pop}}$	$\hat{\mathbb{V}}_{\text{cond}}$	$\hat{\mathbb{V}}_{\text{pop}}$	$\hat{\mathbb{V}}_{\text{cond}}$	$\sqrt{\hat{\mathbb{V}}_{\text{pop}}}$	$\sqrt{\hat{\mathbb{V}}_{\text{cond}}}$
	Misspec	Homo	Samp Size	High Lev	K						
I	No	Yes	50	No	1	0.946	0.941	0.946	0.941	0.378	0.387
II	No	Yes	50	No	5	0.934	0.929	0.934	0.929	0.419	0.428
III	No	Yes	50	Yes	1	0.946	0.940	0.946	0.940	0.370	0.379
IV	No	Yes	50	Yes	5	0.933	0.927	0.933	0.927	0.412	0.420
V	No	Yes	200	No	1	0.941	0.934	0.941	0.934	0.286	0.291
VI	No	Yes	200	No	5	0.947	0.945	0.947	0.945	0.191	0.191
VII	No	Yes	200	Yes	1	0.948	0.945	0.948	0.945	0.183	0.183
VIII	No	Yes	200	Yes	5	0.947	0.945	0.947	0.945	0.188	0.187
IX	No	No	50	No	1	0.947	0.941	0.948	0.941	0.350	0.358
X	No	No	50	No	5	0.935	0.930	0.936	0.931	0.383	0.389
XI	No	No	50	Yes	1	0.943	0.935	0.945	0.937	0.340	0.346
XII	No	No	50	Yes	5	0.930	0.924	0.933	0.927	0.371	0.377

Continued on next page

Table 3.4: *(continued)*

Estimand ⟶						θ_{pop}		θ_{cond}		median	
Variance ⟶						$\hat{\mathbb{V}}_{\text{pop}}$	$\hat{\mathbb{V}}_{\text{cond}}$	$\hat{\mathbb{V}}_{\text{pop}}$	$\hat{\mathbb{V}}_{\text{cond}}$	$\sqrt{\hat{\mathbb{V}}_{\text{pop}}}$	$\sqrt{\hat{\mathbb{V}}_{\text{cond}}}$
	Misspec	Homo	Samp Size	High Lev	K						
XIII	No	No	200	No	1	0.950	0.946	0.950	0.946	0.173	0.173
XIV	No	No	200	No	5	0.945	0.941	0.945	0.941	0.177	0.176
XV	No	No	200	Yes	1	0.946	0.942	0.947	0.942	0.170	0.169
XVI	No	No	200	Yes	5	0.944	0.940	0.945	0.940	0.173	0.172
XVII	Yes	Yes	50	No	1	0.965	0.649	0.999	0.926	0.114	0.055
XVIII	Yes	Yes	50	No	5	0.957	0.802	0.999	0.978	0.136	0.089
XIX	Yes	Yes	50	Yes	1	0.962	0.658	0.999	0.926	0.120	0.059
XX	Yes	Yes	50	Yes	5	0.930	0.872	0.978	0.948	0.399	0.333
XXI	Yes	Yes	200	No	1	0.955	0.636	1.000	0.942	0.056	0.026
XXII	Yes	Yes	200	No	5	0.954	0.715	1.000	0.970	0.058	0.032
XXIII	Yes	Yes	200	Yes	1	0.955	0.648	1.000	0.941	0.058	0.028
XXIV	Yes	Yes	200	Yes	5	0.953	0.724	1.000	0.972	0.061	0.034

Continued on next page

Table 3.4: *(continued)*

Estimand ⟶						θ_{pop}		θ_{cond}		median	
Variance ⟶						$\hat{\mathbb{V}}_{\text{pop}}$	$\hat{\mathbb{V}}_{\text{cond}}$	$\hat{\mathbb{V}}_{\text{pop}}$	$\hat{\mathbb{V}}_{\text{cond}}$	$\sqrt{\hat{\mathbb{V}}_{\text{pop}}}$	$\sqrt{\hat{\mathbb{V}}_{\text{cond}}}$
	Misspec	Homo	Samp Size	High Lev	K						
XXV	Yes	No	50	No	1	0.963	0.635	1.000	0.924	0.120	0.056
XXVI	Yes	No	50	No	5	0.956	0.784	0.999	0.980	0.138	0.089
XXVII	Yes	No	50	Yes	1	0.960	0.764	1.000	0.924	0.127	0.061
XXVIII	Yes	No	50	Yes	5	0.956	0.795	0.999	0.978	0.145	0.096
XXIX	Yes	No	200	No	1	0.956	0.627	1.000	0.942	0.058	0.027
XXX	Yes	No	200	No	5	0.954	0.702	1.000	0.971	0.061	0.033
XXXI	Yes	No	200	Yes	1	0.942	0.887	0.978	0.943	0.356	0.301
XXXII	Yes	No	200	Yes	5	0.953	0.710	1.000	0.971	0.064	0.035

3.7 Conclusion

In this paper we discuss inference for conditional estimands in misspecified models. Following the work by Eicker (1967), Huber (1967), and White (1980a,b, 1982), it is common in empirical work to report robust standard errors. These robust standard errors are valid for the population value of the estimator given random sampling. We show that if one is interested in the conditional estimand, conditional on all or a subset of the variables, robust standard errors are generally smaller than the White robust standard errors. We derive a general characterization of the variance for the conditional estimand and propose a consistent estimator for this variance. We argue that in some settings the conditional estimand may be of more interest than the unconditional one.

References

Abadie, A., Athey, S., Imbens, G. W. and Wooldridge, J. M. (2014). *Finite population causal standard errors* (NBER Working Paper 20325). Cambridge, MA: National Bureau of Economic Research.

— and Imbens, G. W. (2006). Large sample properties of matching estimators for average treatment effects. *Econometrica*, **74** (1), 235–267.

— and — (2008a). Estimation of the conditional variance in paired experiments. *Annales d'Economie et de Statistique*, **91**, 175–187.

— and — (2008b). On the failure of the bootstrap for matching estimators. *Econometrica*, **76** (6), 1537–1557.

Anderson, T. W. and Rubin, H. (1949). Estimation of the parameters of a single equation in a complete system of stochastic equations. *The Annals of Mathematical Statistics*, **20** (1), 46–63.

Andrews, D. W. (1994). Empirical process methods in econometrics. In D. McFadden and R. Engle (eds.), *Handbook of econometrics*, vol. 4, Amsterdam: North Holland.

Angrist, J., Chernozhukov, V. and Fernández-Val, I. (2006). Quantile regression under misspecification, with an application to the us wage structure. *Econometrica*, **74** (2), 539–563.

Angrist, J. D. and Pischke, J.-S. (2008). *Mostly harmless econometrics: An empiricist's companion*. Princeton university press.

Bajari, P., Benkard, C. L. and Levin, J. (2007). Estimating dynamic models of imperfect competition. *Econometrica*, **75** (5), p. 1331–1370.

Basker, E. (2007). The causes and consequences of Wal-Mart's growth. *The Journal of Economic Perspectives*, pp. p. 177–198.

Berry, S. (1992). Estimation of a model of entry in the airline industry. *Econometrica*, pp. p. 889–917.

—, Levinsohn, J. and Pakes, A. (1995). Automobile prices in market equilibrium. *Econometrica*, pp. 841–890.

Bradley, S., Ghemawat, P. and Foley, S. (2002). Wal-Mart stores, inc.

Bresnahan, T. and Reiss, P. (1991). Entry and competition in concentrated markets. *Journal of Political Economy*, pp. p. 977–1009.

Chow, G. C. (1984). Maximum-likelihood estimation of misspecified models. *Economic Modelling*, **1** (2), 134–138.

Efron, B. (1982). *The jackknife, the bootstrap and other resampling plans*, vol. 38. SIAM.

— and Tibshirani, R. J. (1994). *An introduction to the bootstrap*. CRC press.

Eicker, F. (1967). Limit theorems for regressions with unequal and dependent errors. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, vol. 1, pp. 59–82.

Ellickson, P., Houghton, S. and Timmins, C. (2013). Estimating network economies in retail chains: A revealed preference approach. *The RAND Journal of Economics*, **44** (2), p. 169–193.

— and Misra, S. (2008). Supermarket pricing strategies. *Marketing Science*, **27** (5), p. 811–828.

Fortunato, S. and Castellano, C. (2012). Community structure in graphs. In *Computational Complexity*, Springer, pp. p. 490–512.

Fox, E., Montgomery, A. and Lodish, L. (2004). Consumer shopping and spending across retail formats. *The Journal of Business*, **77** (S2), p. 25–60.

Gelman, A. and Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

Goldberger, A. S. (1991). *A course in econometrics*. Harvard University Press.

Gorman, W. (1959). Separable utility and aggregation. *Econometrica*, pp. p. 469–481.

— (1971). Two-stage budgeting. *Unpublished Paper - L.S.E. Working Paper*.

Holmes, T. (2011). The diffusion of Wal-Mart and economies of density. *Econometrica*, **79** (1), p. 253–302.

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, vol. 1, pp. 221–233.

Igami, M. and Yang, N. (2014). Cannibalization and preemptive entry in heterogeneous markets. *Unpublished Paper - Yale Working Paper*.

Imbens, G., Spady, R. H. and Johnson, P. (1998). Information theoretic approaches to inference in moment condition models. *Econometrica*, **66** (2), 333–358.

Imbens, G. W. (1997). One-step estimators for over-identified generalized method of moments models. *The Review of Economic Studies*, **64** (3), 359–383.

— and Kolesar, M. (2012). *Robust standard errors in small samples: Some practical advice* (NBER Working Paper 18478). Cambridge, MA: National Bureau of Economic Research.

—, Rubin, D. B. and Sacerdote, B. I. (2001). Estimating the effect of unearned income on labor earnings, savings, and consumption: Evidence from a survey of lottery players. *The American Economic Review*, **91** (4), 778–794.

Jia, P. (2008). What Happens When Wal-Mart comes to town: An empirical analysis of the discount retailing industry. *Econometrica*, **76** (6), p. 1263–1316.

Koenker, R. and Bassett Jr, G. (1978). Regression quantiles. *Econometrica*, pp. 33–50.

Mazzeo, M. J. (2002). Product choice and oligopoly market structure. *The RAND Journal of Economics*, pp. p. 221–242.

Miller, G. L., Teng, S.-H., Thurston, W. and Vavasis, S. A. (1997). Separators for sphere-packings and nearest neighbor graphs. *Journal of the ACM (JACM)*, **44** (1), 1–29.

Müller, U. K. (2013). Risk of bayesian inference in misspecified models, and the sandwich covariance matrix. *Econometrica*, **81** (5), 1805–1849.

Neumark, D., Zhang, J. and Ciccarella, S. (2008). The effects of Wal-Mart on local labor markets. *Journal of Urban Economics*, **63** (2), 405–430.

Newey, W. K. and McFadden, D. (1994). Estimation in large samples. In D. McFadden and R. Engle (eds.), *Handbook of Econometrics*, vol. 4, Amsterdam: North Holland.

— and Smith, R. J. (2004). Higher order properties of gmm and generalized empirical likelihood estimators. *Econometrica*, **72** (1), 219–255.

Pakes, A. (1986). Patents as options: some estimates of the value of holding european patent stocks. *Econometrica*, **54** (4), 755–784.

—, Ostrovsky, M. and Berry, S. (2007). Simple estimators for the parameters of discrete dynamic games (with entry/exit examples). *The RAND Journal of Economics*, **38** (2), 373–399.

— and Pollard, D. (1989). Simulation and the asymptotics of optimization estimators. *Econometrica*, pp. 1027–1057.

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, **2** (11), 559–572.

Pencavel, J. (1986). Labor supply of men: A survey. In O. Ashenfelter and R. Layard (eds.), *Handbook of Labor Economics*, vol. 1, Amsterdam: North Holland.

Powell, J. L. (1984). Least absolute deviations estimation for the censored regression model. *Journal of Econometrics*, **25** (3), 303–325.

Qin, J. and Lawless, J. (1994). Empirical likelihood and general estimating equations. *The Annals of Statistics*, **22** (1), 300–325.

Reinganum, J. (1981). Market structure and the diffusion of new technology. *The Bell Journal of Economics*, pp. p. 618–624.

Rust, J. (1987). Optimal replacement of GMC bus engines: An empirical model of Harold Zurcher. *Econometrica*, pp. p. 999–1033.

Ryan, S. P. (2012). The costs of environmental regulation in a concentrated industry. *Econometrica*, **80** (3), 1019–1061.

Sachs, J. D. and Warner, A. M. (1997). Fundamental sources of long-run growth. *The American Economic Review*, **87** (2), 184–188.

Schmidt-Dengler, P. (2006). The timing of new technology adoption: The case of MRI. *Unpublished Paper - L.S.E. Working Paper*.

Seim, K. (2006). An empirical model of firm entry with endogenous product-type choices. *The RAND Journal of Economics*, **37** (3), p. 619–640.

Shoag, D. and Veuger, S. (2014). Shops and the city: evidence on local externalities and local government policy from big box bankruptcies. *Unpublished Paper - Harvard Kennedy School Working Paper*.

Spearman, C. (1904). "General intelligence," objectively determined and measured. *The American Journal of Psychology*, **15** (2), 201–292.

Tibshirani, R. (1986). Discussion: Jackknife, bootstrap and other resampling methods in regression analysis. *The Annals of Statistics*, **14** (4), 1335–1339.

Van der Vaart, A. W. (2000). *Asymptotic statistics*. Cambridge: Cambridge University Press.

Van Der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence*. New York: Springer.

White, H. (1980a). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, **48** (4), 817–838.

— (1980b). Using least squares to approximate unknown regression functions. *International Economic Review*, **21** (1), 149–170.

— (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, **50** (1), 1–25.

Wooldridge, J. M. (2002). *Econometric analysis of cross section and panel data*. Cambridge, MA: MIT Press.

Wu, C.-F. J. (1986). Jackknife, bootstrap and other resampling methods in regression analysis. *the Annals of Statistics*, **14** (4), 1261–1295.

Yatchew, A. (1997). An elementary estimator of the partial linear model. *Economics letters*, **57** (2), 135–143.

— (1999). An elementary nonparametric differencing test of equality of regression functions. *Economics Letters*, **62** (3), 271–278.

Zhu, T. and Singh, V. (2009). Spatial competition with endogenous location choices: An application to discount retailing. *Quantitative Marketing and Economics*, **7** (1), p. 1–35.

Appendix A

Appendix to Chapter 1

A.1 Proofs

Proof of Theorem 1.1: Suppose $\exists \sigma_j' \neq \sigma_j^*$ s.t. $\pi(\sigma_j'(s), \sigma_{-j}^*(s), s) > \pi(\sigma_j^*(s), \sigma_{-j}^*(s), s)$. Since σ_{-j}^* and σ_j^* are separable, $\exists m \in \{1, ..., M\}$, s.t. $\sigma_{jm}^*(s_m, \sigma_{-jm}^*(s_m), B_m^*) \neq \sigma_j'^m(s, \sigma_{-jm}^*(s_m), B)$, and $\sum_{l \in P_m} \pi_{jl}(\sigma_j'^m(s), \sigma_{-jm}^*(s_m), s) > \sum_{l \in P_m} \pi_{jl}(\sigma_j^*(s_m), \sigma_{-jm}^*(s_m), s_m)$. But σ_{jm}^* is the best response of σ_{-jm}^*, and $\Delta\pi(s_j, s_{-j}, l)/\Delta\pi(s_j, s_{-j}, h)$ does not depend on (s_{jn}, s_{-jn}), $\forall l, h \in P_m$ and $m \neq n$. Thus there's no profitable deviation by including s_{jn}, $\forall n \neq m$,

i.e. $\sum_{l \in P_m} \pi_{jl}(\sigma_j'^m(s), \sigma_{-jm}^*(s_m), s) \leq \sum_{l \in P_m} \pi_{jl}(\sigma_j^*(s_m), \sigma_{-jm}^*(s_m), s_m)$.

Proof of Theorem1.2: If all conditions are satisfied, $\pi(s)$ is additively separably in $\{P_1, \cdots, P_M\}$. By results in Gorman (1959), $\{1, \cdots, L\}$ is separably in $\{P_1, \cdots, P_M\}$.

Definition A.1 *Locations* $\{1, \cdots, L\}$*are* <u>separable</u> *in the partition* $\{P_1, \cdots, P_M\}$ *if*

$$\frac{\Delta \mathbb{E}V(s_j, s_{-j}, l)}{\Delta \mathbb{E}V(s_j, s_{-j}, h)} \perp (s_j^k, s_{-j}^k) | \{B_{jm}, B_{-jm}\}_{m=1}^M, \forall l, h \in P_m, \forall k \notin P_n, m \neq n,$$

where $l, h \in P_{m_l}$*, and* $k \in P_{m_k}$*, and*

$$\Delta \mathbb{E}V(s_j, s_{-j}, l) = \mathbb{E}V(s_j^l = 1, s_j^{-l}, s_{-j}) - \mathbb{E}V(s_j^l = 0, s_j^{-l}, s_{-j}).$$

Definition A.2 *Firm j's strategy* σ_j^* *is* <u>separable</u> *in the partition* $\{P_1, \cdots, P_M\}$ *if for given* σ_{-j}*,*

$\exists \sigma_{j1}^*, \sigma_{j2}^*, \cdots \sigma_{jM}^*$ *s.t.*

$$\sigma_{jm}^*(s_{jm}, s_{-jm}, B_m) = \sigma_j^{*m}(s_j, s_{-j}, B),$$

where $\sigma_j^* = \left(\sigma_j^{*1}, \cdots, \sigma_j^{*M}\right)$, $B_m = (B_{jm}^*, B_{-jm})$, $B_{jm}^* = \sum_{l \in P_m} a_j^{*l}$, $B_{-jm} = \sum_{l \in P_m} a_{-j}^l$, $\forall m = 1, \cdots, M$, *and* $\sum_{m=1}^M B_m = B$.

Theorem A.1 *If locations* $\{1, \cdots, L\}$ *are separable in the partition* $\{P_1, \cdots, P_M\}$, *there exists a separable equilibrium.*

Proof of Theorem A.1: Results follow the proof of Theorem 1.1.

Theorem A.2 *The location* $\{1, \cdots, L\}$ *is separable in partition* $\{P_1, \cdots, P_M\}$ *if the value function* $\mathbb{E}V(\cdot)$ *satisfies the following conditions,*

1. $R_j^l(s)$ *is additively separable in partition* $\{P_1, \cdots, P_M\}$,
2. *Distribution cost and fixed cost at location* l *is independent of* z_j^k *and* x_j^k, *where* $k \in P_n$,$m \neq n$,
3. η_j^l *are independently distributed across markets.*

Proof of Theorem A.2: Prove by induction. Rewrite the value function as

$$\mathbb{E}V(s_{it}, s_{-jt}) = \sum_{\tau=t}^{\infty} \beta^{\tau-t} \left(\sum_{s_\tau} \mathbb{E}\pi(s_\tau) P(s_\tau | s_t) \right) = \sum_{\tau=t}^{\infty} \beta^{\tau-t} \mathbb{E}V_\tau(s_t).$$

The first term in the outer sum, $\mathbb{E}V_\tau(s_t) = \mathbb{E}\pi(s_t)$ when $\tau = t$, is separable in $\{P_1, \cdots, P_M\}$ by Theorem 1.2. Assume $\mathbb{E}V_\tau(s_t)$ is separable in $\{P_1, \cdots, P_M\}$ for $\tau = T$, then $\sum_{s_T} \mathbb{E}\pi(s_T) P(s_T | s_t)$ is separable. Apply two-stage budgeting, $P(s_T^l | s_t, \{B_{mt}\}_t^T)$ does not depend on (s_{jt}^k, s_{-jt}^k), $\forall l \in B_m$, $k \in B_n$, $m \neq n$. It is left to show $\mathbb{E}V_{T+1}(s_t)$ is separable.

$$\mathbb{E}V_{T+1}(s_t) = \sum_{s_{T+1}} \mathbb{E}\pi(s_{T+1}) P(s_{T+1} | s_t)$$

$$= \sum_{s_{T+1}} \sum_{s_T} \mathbb{E}\pi(s_{T+1}) P(s_{T+1} | s_T) P(s_T | s_t).$$

Then

$$\Delta \mathbb{E}V_{T+1}(s_{jt}, s_{-jt}, l_{T+1})$$

$$= \sum_{s_T} \sum_{s_{-jT+1}} \Bigg[\left[\mathbb{E}\pi(s_{jT} + l_{T+1}, s_{-jT+1}) - \mathbb{E}\pi(s_{jT}, s_{-jT}) \right] P(s_{-jT+1}|s_T) \Bigg] P(s_T|s_t),$$

where l_{T+1} indicates new store l opened in period $T+1$, and $l \in P_m$. $\mathbb{E}\pi(s_{jT} + l_{T+1}, s_{-jT+1})$ and $\mathbb{E}\pi(s_{jT}, s_{-jT})$ are separable by the separability of the static profit. As a result,

$$\mathbb{E}\pi(s_{jT} + l_{T+1}, s_{-jT+1}) - \mathbb{E}\pi(s_{jT}, s_{-jT}) = \mathbb{E}\pi(s_{jmT} + l_{T+1}, s_{-jmT+1}) - \mathbb{E}\pi(s_{jmT}, s_{-jmT}),$$

where $s_{jmT} = \{s^h_{jT} | h \in P_m\}$, and $s_{-jmT} = \{s^h_{-jT} | h \in P_m\}$. Note

$$P(s_{-jT+1}|s_T, B_{mT+1})$$

$$= P\Bigg[\mathbb{E}\pi(s_{jT+1}, s_{-jT+1}) - \mathbb{E}\pi(s_{jT+1}, s_{-jT}) \geq max_{s'_{-jT+1}} \left(\mathbb{E}\pi(s_{jT+1}, s'_{-jT+1}) - \mathbb{E}\pi(s_{jT+1}, s'_{-jT}) \right) | B_{mT+1} \Bigg],$$

where $s_{jT+1} = s_{jT} + l_{T+1}$, and $s'_{-jT+1} \in \{s_{-jT+1} | s_{-jT+1} = s_{-jT} + h_{T+1}, h \in P_m\}$,

$$= P\Bigg[\mathbb{E}\pi(s_{jmT+1}, s_{-jmT+1}) - \mathbb{E}\pi(s_{jmT+1}, s_{-jmT}) \geq max_{s'_{-jmT+1}} \left(\mathbb{E}\pi(s_{jmT+1}, s_{-jmT+1}) - \mathbb{E}\pi(s_{jmT+1}, s'_{-jmT}) \right) \Bigg],$$

where $s_{jmT+1} = s_{jmT} + l_{T+1}$. Thus $\sum_{s_{-jT+1}} \left[\left[\mathbb{E}\pi(s_{jT} + l_{T+1}, s_{-jT+1}) - \mathbb{E}\pi(s_{jT}, s_{-jT}) \right] P(s_{-jT+1}|s_T) \right]$ is additively separable in $\{P_1, \cdots, P_M\}$. Since $\mathbb{E}V_T(s_t)$ is separable,

$$\frac{\Delta \mathbb{E}V_{T+1}(s_{jt}, s_{-jt}, l_{T+1})}{\Delta \mathbb{E}V_{T+1}(s_{jt}, s_{-jt}, h_{T+1})} \perp (s^k_{jt}, s^k_{-jt}) | \{(B_{jmt}, B_{-jmt})_{t=1}^{T+1}\},$$

$\forall l, h \in P_m$, and $k \in P_n$, $m \neq n$.

Appendix B

Appendix to Chapter 2

B.1 Simulation method for computing standard errors

In this section, I describe how the first stage estimation error and second stage clustering error can be accounted for in the standard errors of the structural estimate in the third stage. It is a simulation method and has four steps.

1. Denote $\hat{\beta}$ and $f(\hat{\beta})$ the estimated demand parameter and its distribution from the first stage estimation. Take R draws of $\hat{\beta}$, $\{\hat{\beta}^r\}_{r=1}^R$, from $f(\hat{\beta})$.

2. Recompute market divisions using the clustering algorithm described in Section 1.3.3.5 for a given $\hat{\beta}^r$. Denote the market divisions by $\{P_1^r, \cdots P_M^r\}$.

3. Given demand estimate $\hat{\beta}^r$ and market division $\{P_1^r, \cdots P_M^r\}$, estimate the structural model to get $(\hat{\psi}^r, \hat{\alpha}^r)$ and its distribution $f(\hat{\psi}^r, \hat{\alpha}^r)$.

4. Repeat the previous two steps R times and compare $f(\hat{\psi}^r, \hat{\alpha}^r)$ to see if the first stage and second stage errors have an impact on the standard errors of $(\hat{\psi}^r, \hat{\alpha}^r)$.

Note the clustering error in the second stage is treated as a machine error. To properly account for the clustering error, one would need model store locations on random field. This is an interesting direction for future research.

Appendix C

Appendix to Chapter 3

C.1 Proofs

Proof of Theorem 3.1: Given Assumptions 3.1 and 3.2, Theorem 2.6 in Newey and McFadden (1994) implies the first result. To prove the second result, let $\rho(x,\theta) = \mathbb{E}[\psi(Y_i, X_i, \theta)|X_i = x]$. Notice that $\mathbb{E}[\rho(X_i, \theta_{\text{pop}})] = 0$. Therefore, $\theta_{\text{cond}}(\mathbf{X})$ can be thought of as an extremum estimator that minimizes

$$\left(\frac{1}{N}\sum_{i=1}^{N}\rho(X_i,\theta)\right)'\left(\frac{1}{N}\sum_{i=1}^{N}\rho(X_i,\theta)\right).$$

We will prove $\theta_{\text{cond}}(\mathbf{X}) - \theta_{\text{pop}} \xrightarrow{p} 0$ by showing that Assumption 3.2 also holds if we replace $\psi(Y_i, X_i, \theta)$ with $\rho(X_i, \theta)$. Because $\mathbb{E}[\rho(X_i,\theta)] = \mathbb{E}[\psi(Y_i, X_i, \theta)]$ it follows that part *(i)* in Assumption 2 holds also with $\rho(X_i, \theta)$ replacing $\psi(Y_i, X_i, \theta)$. Part *(ii)* of Assumption 2 follows from dominated convergence because, by Assumption 2*(iii)*, $\mathbb{E}[\sup_{\theta\in\Theta}\|\psi(Y_i, X_i, \theta)\||X_i] < \infty$ with probability one. To prove that part *(iii)* holds also after replacing $\psi(Y_i, X_i, \theta)$ with $\rho(X_i, \theta)$, notice that,

$$\|\rho(X_i,\theta)\| = \|\mathbb{E}[\psi(Y_i, X_i, \theta)|X_i]\| \leq \mathbb{E}\left[\|\psi(Y_i, X_i, \theta)\|\,\middle|\,X_i\right],$$

because the norm is a convex function by the Triangle Inequality. Therefore,

$$\sup_{\theta\in\Theta}\|\rho(X_i,\theta)\| \leq \sup_{\theta\in\Theta}\mathbb{E}\left[\|\psi(Y_i,X_i,\theta)\|\,\middle|X_i\right] \leq \mathbb{E}\Big[\sup_{\theta\in\Theta}\|\psi(Y_i,X_i,\theta)\|\,\Big|X_i\Big].$$

Taking expectations on both sides of the previous equation and using Assumption 3.2*(iii)*, we obtain $\mathbb{E}[\sup_{\theta\in\Theta}\|\rho(X_i,\theta)\|] < \infty$. Now, Theorem 2.6 in Newey and McFadden (1994) implies $\theta_{\text{cond}}(\mathbf{X}) - \theta_{\text{pop}} \xrightarrow{p} 0$ and, therefore, the second result of the theorem. □

Proof of Theorem 3.2: The first result follows from Theorem 3.4 in Newey and McFadden (1994).

To prove the second result, we will first establish the joint asymptotic distribution of $\sqrt{N}(\hat{\theta}-\theta_{\text{pop}})$ and $\sqrt{N}(\theta_{\text{cond}}(\mathbf{X})-\theta_{\text{pop}})$, and then we use this result to derive the asymptotic distribution of

$$\sqrt{N}(\hat{\theta}-\theta_{\text{cond}}(\mathbf{X})) = \sqrt{N}(\hat{\theta}-\theta_{\text{pop}}) - \sqrt{N}(\theta_{\text{cond}}(\mathbf{X})-\theta_{\text{pop}}). \tag{C.1.1}$$

By Assumption 3.3*(ii)* and *(iv)* and Lemma 3.6 in Newey and McFadden (1994) we obtain that, for x in a set of probability one, $\rho(x,\theta)$ is continuously differentiable with respect to θ in an open neighborhood $\mathcal{N}$ of θ_{pop}, with

$$\frac{\partial\rho(x,\theta)}{\partial\theta'} = \mathbb{E}\left[\frac{\partial\psi(Y_i,X_i,\theta)}{\partial\theta'}\,\middle|\,X_i=x\right].$$

Notice that

$$\begin{aligned}\psi(Y_i,X_i,\theta_{\text{pop}})'\psi(Y_i,X_i,\theta_{\text{pop}}) &= \mathbb{E}[\psi(Y_i,X_i,\theta_{\text{pop}})'|X_i]\mathbb{E}[\psi(Y_i,X_i,\theta_{\text{pop}})|X_i]\\ &+ (\psi(Y_i,X_i,\theta_{\text{pop}}) - \mathbb{E}[\psi(Y_i,X_i,\theta_{\text{pop}})|X_i])'(\psi(Y_i,X_i,\theta_{\text{pop}}) - \mathbb{E}[\psi(Y_i,X_i,\theta_{\text{pop}})|X_i])\\ &+ 2E[\psi(Y_i,X_i,\theta_{\text{pop}})'|X_i](\psi(Y_i,X_i,\theta_{\text{pop}}) - \mathbb{E}[\psi(Y_i,X_i,\theta_{\text{pop}})|X_i]).\end{aligned}$$

Taking expectation eliminates the cross-product term, which implies:

$$\mathbb{E}\left[\|\rho(X_i,\theta_{\text{pop}})\|^2\right] \leq \mathbb{E}\left[\|\psi(Y_i,X_i,\theta_{\text{pop}})\|^2\right] < \infty.$$

Using convexity of the norm, we obtain

$$\sup_{\theta\in\mathcal{N}}\left\|\frac{\partial\rho(x,\theta)}{\partial\theta'}\right\| \leq \mathbb{E}\left[\sup_{\theta\in\mathcal{N}}\left\|\frac{\partial\psi(Y_i,X_i,\theta)}{\partial\theta'}\right\| \;\middle|\; X_i = x\right].$$

Taking averages on both sides of the last equation and using Assumption 3.3*(iv)* we obtain:

$$\mathbb{E}\left[\sup_{\theta\in\mathcal{N}}\left\|\frac{\partial\rho(x,\theta)}{\partial\theta'}\right\|\right] < \infty.$$

Notice also that

$$\mathbb{E}\left[\frac{\partial\rho(X_i,\theta_{\text{pop}})}{\partial\theta'}\right] = \mathbb{E}\left[\frac{\partial\psi(Y_i,X_i,\theta_{\text{pop}})}{\partial\theta'}\right] = \Gamma,$$

which is nonsingular by Assumption 3.3*(v)*.

As a result, Theorem 3.4 in Newey and McFadden (1994) holds for the estimator that minimizes

$$\left(\frac{1}{N}\sum_{i=1}^{N}\begin{pmatrix}\psi(Y_i,X_i,\theta_1)\\ \rho(X_i,\theta_2)\end{pmatrix}\right)'\left(\frac{1}{N}\sum_{i=1}^{N}\begin{pmatrix}\psi(Y_i,X_i,\theta_1)\\ \rho(X_i,\theta_2)\end{pmatrix}\right)$$

with respect to θ_1 and θ_2. Applying Theorem 3.4 of Newey and McFadden (1994), we obtain

$$\sqrt{N}\begin{pmatrix}\hat{\theta}-\theta_{\text{pop}}\\ \theta_{\text{cond}}(\mathbf{X})-\theta_{\text{pop}}\end{pmatrix} \xrightarrow{d} N(0,\Gamma_{\text{joint}}^{-1}\mathbb{V}_{\text{joint}}(\Gamma_{\text{joint}}^{-1})'),$$

where $\mathbb{V}_{\text{joint}}$ is equal to

$$\begin{pmatrix}\mathbb{E}\left[\psi(Y_i,X_i,\theta_{\text{pop}})\psi(Y_i,X_i,\theta_{\text{pop}})'\right] & \mathbb{E}\left[\mathbb{E}[\psi(Y_i,X_i,\theta_{\text{pop}})|X_i]\mathbb{E}[\psi(Y_i,X_i,\theta_{\text{pop}})'|X_i]\right]\\ \mathbb{E}\left[\mathbb{E}[\psi(Y_i,X_i,\theta_{\text{pop}})|X_i]\mathbb{E}[\psi(Y_i,X_i,\theta_{\text{pop}})'|X_i]\right] & \mathbb{E}\left[\mathbb{E}[\psi(Y_i,X_i,\theta_{\text{pop}})|X_i]\mathbb{E}[\psi(Y_i,X_i,\theta_{\text{pop}})'|X_i]\right]\end{pmatrix},$$

and

$$\Gamma_{\text{joint}} = \begin{pmatrix}\Gamma & 0\\ 0 & \Gamma\end{pmatrix}.$$

Now, because equation (C.1.1), we obtain,

$$\sqrt{N}\left(\hat{\theta}-\theta_{\text{cond}}(\mathbf{X})\right) \xrightarrow{d} N(0,\mathbb{V}_{\text{gmm,cond}}),$$

where $\mathbb{V}_{\text{gmm,cond}} = \Gamma^{-1}\Delta_{\text{cond}}(\Gamma^{-1})'$, and

$$\begin{aligned}\Delta_{\text{cond}} &= \mathbb{E}\left[\psi(Y_i, X_i, \theta_{\text{pop}})\psi(Y_i, X_i, \theta_{\text{pop}})'\right] - \mathbb{E}\left[\mathbb{E}[\psi(Y_i, X_i, \theta_{\text{pop}})|X_i]\mathbb{E}[\psi(Y_i, X_i, \theta_{\text{pop}})'|X_i]\right] \\ &= \mathbb{E}\left[\mathbb{V}(\psi(Y_i, X_i, \theta_{\text{pop}})|X_i)\right].\end{aligned}$$

□

Proof of Corollary 3.1: The result follows directly from $\mathbb{V}(\psi(Y_i, X_i, \theta_{\text{pop}})|X_i) = \mathbb{E}[\psi(Y_i, X_i, \theta_{\text{pop}})\psi(Y_i, X_i, \theta_{\text{pop}})'|X_i] - \mathbb{E}[\psi(Y_i, X_i, \theta_{\text{pop}})|X_i]\mathbb{E}[\psi(Y_i, X_i, \theta_{\text{pop}})'|X_i]$. □

We next state a lemma from Abadie and Imbens (2010) that will be useful in what follows.

Lemma C.1 (Lemma 1, Abadie and Imbens (2010, page 180)) *Suppose that $W_1, W_2, \ldots$ is a sequence with $W_i \in \mathbb{W}$ where $\mathbb{W}$ a compact subset of $\mathbb{R}^K$. Then*

$$\lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} \left\| W_i - W_{\ell_W(i)} \right\|^2 = 0.$$

Lemma C.2 (Average Conditional Moments) *Let (V_i, W_i), $i = 1, \ldots, N$, be a sequence of independent, identically distributed random variables, with V_i scalar, and with compact support for W_i. For some positive integer n, and for $j = 1, 2, \ldots, n$, let $\mu_p(w) = \mathbb{E}[V_i^p|W_i = w]$ be Lipschitz in w with constant C_p. Then for all nonnegative k, m such that $\max(k, m) \leq n/2$,*

$$\frac{1}{N}\sum_{i=1}^{N} V_i^k \cdot V_{\ell_W(i)}^m \xrightarrow{p} \mathbb{E}\left[\mathbb{E}\left(V_i^k \middle| W_i\right) \cdot \mathbb{E}\left(V_i^m \middle| W_i\right)\right].$$

Proof of Lemma C.2: First we show

$$\mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N} V_i^k \cdot V_{\ell_W(i)}^m - \mathbb{E}\left[\mathbb{E}\left(V_i^k \middle| W_i\right) \cdot \mathbb{E}\left(V_i^m \middle| W_i\right)\right]\right] = o(1). \tag{C.1.2}$$

Because V_i and $V_{\ell_W(i)}$ are independent conditional on $\mathbf{W} = (W_1, ..., W_N)'$,

$$\begin{aligned}\mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N} V_i^k \cdot V_{\ell_W(i)}^m\right] &= \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}\left\{\mathbb{E}\left[V_i^k \cdot V_{\ell_W(i)}^m \middle| \mathbf{W}\right]\right\} \\ &= \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}\left\{\mathbb{E}(V_i^k \middle| \mathbf{W}) \cdot \mathbb{E}\left(V_{\ell_W(i)}^m \middle| \mathbf{W}\right)\right\}\end{aligned}$$

$$= \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}\left\{ \mathbb{E}(\, V_i^k \Big|\, W_i) \cdot \mathbb{E}\left(V_{\ell_W(i)}^m \Big|\, W_{\ell_W(i)} \right) \right\}$$

$$= \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}\left[\mu_k(W_i) \cdot \mu_m\left(W_{\ell_W(i)} \right) \right]$$

$$= \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}\left\{ \mu_k(W_i) \cdot \left[\mu_m(W_i) + \mu_m\left(W_{\ell_W(i)} \right) - \mu_m(W_i) \right] \right\}$$

$$= \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}\left[\mu_k(W_i) \cdot \mu_m(W_i) \right] + \mathbb{E}\left\{ \frac{1}{N}\sum_{i=1}^{N} \mu_k(W_i) \left[\mu_m\left(W_{\ell_W(i)} \right) - \mu_m(W_i) \right] \right\}$$

$$= \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}\left[\mathbb{E}\left(V_i^k | W_i \right) \cdot \mathbb{E}\left(V_i^m | W_i \right) \right] + \mathbb{E}\left\{ \frac{1}{N}\sum_{i=1}^{N} \mu_k(W_i) \left[\mu_m\left(W_{\ell_W(i)} \right) - \mu_m(W_i) \right] \right\}.$$

Therefore,

$$\left| \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N} V_i^k \cdot V_{\ell_W(i)}^m - \mathbb{E}\left[\mathbb{E}\left(V_i^k | W_i \right) \cdot \mathbb{E}\left(V_i^m | W_i \right) \right] \right] \right|$$

$$= \left| \mathbb{E}\left\{ \frac{1}{N}\sum_{i=1}^{N} \mu_k(W_i) \left[\mu_m\left(W_{\ell_W(i)} \right) - \mu_m(W_i) \right] \right\} \right|$$

$$\leq \mathbb{E}\left\{ \frac{1}{N}\sum_{i=1}^{N} |\mu_k(W_i)| \cdot \left| \mu_m\left(W_{\ell_W(i)} \right) - \mu_m(W_i) \right| \right\}$$

$$\leq \sup_w |\mu_k(w)| \cdot \mathbb{E}\left\{ \frac{1}{N}\sum_{i=1}^{N} C_m \left\| W_i - W_{\ell_W(i)} \right\| \right\}$$

$$= o(1),$$

by Lemma C.1 and dominated convergence. This finishes the proof of (C.1.2).

Next, we will show that

$$\mathbb{E}\left\{ \left[\frac{1}{N}\sum_{i=1}^{N} V_i^k \cdot V_{\ell_W(i)}^m - \mathbb{E}\left[\mathbb{E}\Big(V_i^k \Big|\, W_i \Big) \cdot \mathbb{E}\Big(V_i^m |\, W_i \Big) \right] \right]^2 \right\} = o(1), \tag{C.1.3}$$

which, together with (C.1.2), proves the claim in the Lemma. First we expand the square:

$$\mathbb{E}\left\{ \left[\frac{1}{N}\sum_{i=1}^{N} V_i^k \cdot V_{\ell_W(i)}^m - \mathbb{E}\left[\mathbb{E}\Big(V_i^k \Big|\, W_i \Big) \cdot \mathbb{E}\Big(V_i^m |\, W_i \Big) \right] \right]^2 \right\}$$

$$= \mathbb{E}\left\{ \left[\frac{1}{N}\sum_{i=1}^{N} V_i^k \cdot V_{\ell_W(i)}^m \right]^2 \right\} + \left\{ \mathbb{E}\left[\mathbb{E}\left(V_i^k | W_i \right) \cdot \mathbb{E}\left(V_i^m | W_i \right) \right] \right\}^2$$

$$-2\mathbb{E}\left\{\frac{1}{N}\sum_{i=1}^{N} V_i^k \cdot V_{\ell_W(i)}^m \cdot \mathbb{E}\left[\mathbb{E}\left(V_i^k|W_i\right)\cdot\mathbb{E}\left(V_i^m|W_i\right)\right]\right\}$$

By (C.1.2), this is equal to

$$\mathbb{E}\left[\left(\frac{1}{N}\sum_{i=1}^{N} V_i^k \cdot V_{\ell_W(i)}^m\right)^2\right] - \left\{\mathbb{E}\left[\mathbb{E}\left(V_i^k|W_i\right)\cdot\mathbb{E}\left(V_i^m|W_i\right)\right]\right\}^2 + o(1)$$

$$= \frac{1}{N^2}\sum_{i=1}^{N}\mathbb{E}\left[V_i^{2k}\cdot V_{\ell_W(i)}^{2m}\right] + \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j\neq i}\mathbb{E}\left[V_i^k V_{\ell_W(i)}^m V_j^k V_{\ell_W(j)}^m\right] \tag{C.1.4}$$

$$- \left\{\mathbb{E}\left[\mathbb{E}\left(V_i^k|W_i\right)\cdot\mathbb{E}\left(V_i^m|W_i\right)\right]\right\}^2 + o(1).$$

Consider the first term in (C.1.4). Using the independence of V_i and $V_{\ell_W(i)}$ conditional on $\mathbf{W}$ we have

$$\frac{1}{N^2}\sum_{i=1}^{N}\mathbb{E}\left[V_i^{2k}\cdot V_{\ell_W(i)}^{2m}\right] = \frac{1}{N^2}\sum_{i=1}^{N}\mathbb{E}\left[\mathbb{E}\left[V_i^{2k}|W_i\right]\cdot\mathbb{E}\left[V_{\ell_W(i)}^{2m}|W_{\ell_W(i)}\right]\right]$$

$$= \frac{1}{N^2}\sum_{i=1}^{N}\mathbb{E}\left[\mu_{2k}(W_i)\cdot\mu_{2m}(W_{\ell_W(i)})\right] \leq \frac{C}{N},$$

because the terms are bounded by the Lipschitz condition on $\mu_p(x)$ for all p at least equal to $2k$ and $2m$. Therefore the first term in (C.1.4) is $o(1)$, and the entire expression is

$$\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j\neq i}\mathbb{E}\left[V_i^k V_{\ell_W(i)}^m V_j^k V_{\ell_W(j)}^m\right] - \left\{\mathbb{E}\left[\mathbb{E}\left(V_i^k|W_i\right)\cdot\mathbb{E}\left(V_i^m|W_i\right)\right]\right\}^2 + o(1). \tag{C.1.5}$$

We write the expectation of the first term conditional on $\mathbf{W}$ as

$$\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j\neq i}\mathbb{E}\left[\mathbb{E}\left[V_i^k V_{\ell_W(i)}^m V_j^k V_{\ell_W(j)}^m\middle|\mathbf{W}\right]\right]$$

$$= \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j\neq i,\ell_W(i)\neq j,\ell_W(j)\neq i}\mathbb{E}\left[\mathbb{E}\left[V_i^k V_{\ell_W(i)}^m V_j^k V_{\ell_W(j)}^m\middle|\mathbf{W}\right]\right]$$

$$+\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j\neq i,\ell_W(i)=j \text{ or } \ell_W(j)=i}\mathbb{E}\left[\mathbb{E}\left[V_i^k V_{\ell_W(i)}^m V_j^k V_{\ell_W(j)}^m\middle|\mathbf{W}\right]\right].$$

The number of terms in the second sum is limited by the "kissing number" the number of units a given unit can be the closest match for (Miller *et al.*, 1997, see also Abadie and Imbens, 2008a), which depends on the dimension of W_i. Let the kissing number be denoted by $\overline{L}$. Then, for given i there is only one j such that $\ell_W(i) = j$, and at most $\overline{L}$ j such that

$\ell_W(j) = i$. With each term in the second sum bounded by $\mathbb{E}[V_i^{m+k}|W_i] \cdot \mathbb{E}[V_i^{m+k}|W_i]$, which is bounded, the second sum is bounded by

$$\mathbb{E}\left[\frac{\overline{L}}{N} \cdot \mathbb{E}[V_i^{m+k}|W_i]^2\right] = o(1).$$

Hence

$$\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j\neq i}\mathbb{E}\left[\mathbb{E}\left[V_i^k V_{\ell_W(i)}^m V_j^k V_{\ell_W(j)}^m \middle| \mathbf{W}\right]\right]$$

$$= \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j\neq i,\ell_W(i)\neq j,\ell_W(j)\neq i}\mathbb{E}\left[\mathbb{E}\left[V_i^k V_{\ell_W(i)}^m V_j^k V_{\ell_W(j)}^m \middle| \mathbf{W}\right]\right] + o(1)$$

$$= \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j\neq i,\ell_W(i)\neq j,\ell_W(j)\neq i}\mathbb{E}\left\{\mathbb{E}\left(V_i^k|W_i\right)\mathbb{E}\left(V_{\ell_W(i)}^m|W_{\ell_W(i)}\right)\mathbb{E}\left(V_j^k|W_j\right)\mathbb{E}\left(V_{\ell_W(j)}^m|W_{\ell_W(j)}\right)\right\} + o(1). \tag{C.1.6}$$

Because of the Lipschitz condition on $\mu_p(w) = \mathbb{E}[V_i^p|W_i = w]$ it follows that

$$\left|\mathbb{E}\left(V_i^k|W_i\right)\mathbb{E}\left(V_{\ell_W(i)}^m|W_{\ell_W(i)}\right)\mathbb{E}\left(V_j^k|W_j\right)\mathbb{E}\left(V_{\ell_W(j)}^m|W_{\ell_W(j)}\right)\right.$$

$$\left.-\mathbb{E}\left(V_i^k|W_i\right)\mathbb{E}\left(V_i^m|W_i\right)\mathbb{E}\left(V_j^k|W_j\right)\mathbb{E}\left(V_j^m|W_j\right)\right|$$

$$\leq C \cdot \max_i \|W_i - W_{\ell_W(i)}\| \cdot \max_j \|W_j - W_{\ell_W(j)}\|,$$

which goes to zero by Lemma A.1. Hence (C.1.6) is

$$\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j\neq i,\ell_W(i)\neq j,\ell_W(j)\neq i}\mathbb{E}\left\{\mathbb{E}\left(V_i^k|W_i\right)\mathbb{E}\left(V_i^m|W_i\right)\mathbb{E}\left(V_j^k|W_j\right)\mathbb{E}\left(V_j^m|W_j\right)\right\} + o(1). \tag{C.1.7}$$

Next we show that this is equal to

$$\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j\neq i}\mathbb{E}\left\{\mathbb{E}\left(V_i^k|W_i\right)\mathbb{E}\left(V_i^m|W_i\right)\mathbb{E}\left(V_j^k|W_j\right)\mathbb{E}\left(V_j^m|W_j\right)\right\} + o(1). \tag{C.1.8}$$

The difference between (C.1.7) and (C.1.8) is

$$\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j|j=i \text{ or } \ell_W(i)=j, \text{ or } \ell_W(j)=i}\mathbb{E}\left\{\mathbb{E}\left(V_i^k|W_i\right)\mathbb{E}\left(V_i^m|W_i\right)\mathbb{E}\left(V_j^k|W_j\right)\mathbb{E}\left(V_j^m|W_j\right)\right\}. \tag{C.1.9}$$

All terms in this sum are bounded by the Lipschitz condition. By the bound on the kissing

number and the boundedness of the expectations, it follows that (C.1.9) is $o(1)$. Next,

$$\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j\neq i}\mathbb{E}\left\{\mathbb{E}\left(V_i^k|W_i\right)\mathbb{E}\left(V_i^m|W_i\right)\mathbb{E}\left(V_j^k|W_j\right)\mathbb{E}\left(V_j^m|W_j\right)\right\}$$

$$=\left\{\mathbb{E}\left[\mathbb{E}\Big(V_i^k|W_i\Big)\cdot\mathbb{E}\Big(V_i^m|W_i\Big)\right]\right\}^2+o(1),$$

and thus

$$\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j\neq i}\mathbb{E}\left[V_i^k V_{\ell_W(i)}^m V_j^k V_{\ell_W(j)}^m\right]-\left\{\mathbb{E}\left[\mathbb{E}\Big(V_i^k|W_i\Big)\cdot\mathbb{E}\Big(V_i^m|W_i\Big)\right]\right\}^2+o(1)=o(1),$$

by (C.1.2). This finishes the proof of (C.1.3), and thus the claim in the lemma. □

Lemma C.3 (Average Conditional Variances) *Let (V_i, W_i), $i = 1,\ldots,N$, be a random sample from the distribution of (V,W) where (V,W) are a pair of random vectors, with compact support for W_i. Suppose that $\mu_p(w) = \mathbb{E}[V_i^p|W_i = w]$ is Lipschitz in w with constant C_p for $p \leq 4$. Define*

$$\widehat{\mathbb{V}}_{\text{cond}} = \frac{1}{2N}\sum_{i=1}^{N}\left(V_i - V_{\ell_W(i)}\right)\left(V_i - V_{\ell_W(i)}\right)'.$$

Then:

$$\widehat{\mathbb{V}}_{\text{cond}} \xrightarrow{p} \mathbb{E}\left[\mathbb{V}(V_i|W_i)\right]. \tag{C.1.10}$$

Proof of Lemma C.3: To prove $\widehat{\mathbb{V}}_{\text{cond}} \xrightarrow{p} \mathbb{E}\left[\mathbb{V}(V_i|W_i)\right]$, we show

$$\mathbb{E}\left\{\widehat{\mathbb{V}}_{\text{cond}} - \mathbb{E}\left[\mathbb{V}(V_i|W_i)\right]\right\}^2 = o(1).$$

Without loss of generality we focus on the case with V scalar:

$$\widehat{\mathbb{V}}_{\text{cond}} = \frac{1}{2N}\sum_{i=1}^{N}\left(V_i - V_{\ell_W(i)}\right)^2 = \frac{1}{2N}\sum_{i=1}^{N}V_i^2 + \frac{1}{2N}\sum_{i=1}^{N}V_{\ell_W(i)}^2 - \frac{1}{N}\sum_{i=1}^{N}V_i V_{\ell_W(i)},$$

and

$$\mathbb{E}\left[\mathbb{V}(V_i|W_i)\right] = \mathbb{E}\left\{\mathbb{E}\left(V_i^2\middle|W_i\right) - \left[\mathbb{E}\left(V_i\middle|W_i\right)\right]^2\right\} = \mathbb{E}\left[V_i^2\right] - \mathbb{E}\left[\mathbb{E}\left(V_i\middle|W_i\right)^2\right].$$

Because $\sum_{i=1}^{N}V_i^2/N \xrightarrow{p} \mathbb{E}[V_i^2]$ by the law of large numbers, it is sufficient to show

$$\frac{1}{N}\sum_{i=1}^{N}V_{\ell_W(i)}^2 \xrightarrow{p} \mathbb{E}\left[V_i^2\right], \qquad \text{and} \quad \frac{1}{N}\sum_{i=1}^{N}V_i\cdot V_{\ell_W(i)} \xrightarrow{p} \mathbb{E}\left[\mathbb{E}\left(V_i\middle|W_i\right)^2\right]. \tag{C.1.11}$$

The first part of (C.1.11) follows from applying Lemma C.2 with $k = 0$ and $m = 2$, and the second part follows from applying Lemma C.2 with $k = m = 1$. □

Proof of Theorem 3.3: Since $\hat{\theta} \xrightarrow{p} \theta_{\text{pop}}$ and $\psi(Y_i, X_i, \theta)$ is differentiable in θ, $\hat{\Gamma} \xrightarrow{p} \Gamma$ by the law of large numbers. Then it is sufficient to show $\hat{\Delta}_{\text{cond}} \xrightarrow{p} \Delta_{\text{cond}}$. Define

$$\tilde{\Delta}_{\text{cond}} = \frac{1}{2N} \sum_{i=1}^{N} \left(\psi(Y_i, X_i, \theta_{\text{cond}}) - \psi(Y_{\ell_X(i)}, X_{\ell_X(i)}, \theta_{\text{cond}}) \right) \left(\psi(Y_i, X_i, \theta_{\text{cond}}) - \psi(Y_{\ell_X(i)}, X_{\ell_X(i)}, \theta_{\text{cond}}) \right)'.$$

Let $V_i = \psi(Y_i, X_i, \theta_{\text{cond}})$, and $W_i = X_i$. By Lemma C.3, $\tilde{\Delta}_{\text{cond}} \xrightarrow{p} \mathbb{V}\left(\psi(Y_i, X_i, \theta_{\text{pop}})\right)$. Because $\hat{\theta} \xrightarrow{p} \theta_{\text{pop}}$ and $\psi(Y_i, X_i, \theta)$ is differentiable in θ, it follows that $\hat{\Delta}_{\text{cond}} - \tilde{\Delta}_{\text{cond}} \xrightarrow{p} 0$. Therefore, $\widehat{\mathbb{V}}_{\text{gmm,cond}} = \hat{\Gamma}^{-1} \hat{\Delta}_{\text{cond}} (\hat{\Gamma}')^{-1} \xrightarrow{p} \Gamma^{-1} \Delta_{\text{cond}} (\Gamma')^{-1} = \mathbb{V}_{\text{gmm,cond}}$. □

C.2 Asymptotic Distribution without Differentiability

Let

$$\hat{g}_N(\theta) = \frac{1}{N} \sum_{i=1}^{N} \psi(Y_i, X_i, \theta),$$
$$g_0(\theta) = \mathbb{E}[\psi(Y_i, X_i, \theta)],$$

and

$$\hat{h}_N(\theta) = \frac{1}{N} \sum_{i=1}^{N} \rho(X_i, \theta).$$

Assumption C.1 *(i)* $\|\hat{g}_N(\hat{\theta})\|^2 \leq \inf_{\theta \in \Theta} \|\hat{g}_N(\theta)\|^2 + o_p(1/N)$ *and* $\|\hat{h}_N(\theta(\mathbf{X}))\|^2 \leq \inf_{\theta \in \Theta} \|\hat{h}_N(\theta)\|^2 + o_p(1/N)$*; (ii)* $g_0(\theta)$ *is differentiable at* θ_{pop} *with non-singular derivative* $\Gamma = \partial g_0(\theta_{\text{pop}})/\partial \theta'$*; (iii)* θ_{pop} *is an interior point of* Θ*; (iv)* $\mathbb{E}[\|\psi(Y_i, X_i, \theta_{\text{pop}})\|^2] < \infty$*; (v) for all* $\delta_N \to 0$, $\sup_{\|\theta - \theta_{\text{pop}}\| \leq \delta_N} \sqrt{N} \|\hat{g}_N(\theta) - \hat{g}_N(\theta_{\text{pop}}) - g_0(\theta)\| \xrightarrow{p} 0$.

Theorem C.1 *Under Assumptions 3.1-3.3 and C.1,*

$$\sqrt{N}(\hat{\theta} - \theta_{\text{pop}}) \xrightarrow{d} \mathcal{N}(0, \Gamma^{-1} \Delta_{\text{pop}} (\Gamma^{-1})'),$$

where $\Delta_{\text{pop}} = \mathbb{E}[\psi(Y_i, X_i, \theta_{\text{pop}}) \psi(Y_i, X_i, \theta_{\text{pop}})']$ *and*

$$\sqrt{N}(\hat{\theta} - \theta_{\text{cond}}(\mathbf{X})) \xrightarrow{d} \mathcal{N}(0, \Gamma^{-1} \Delta_{\text{cond}} (\Gamma^{-1})'),$$

where $\Delta_{\text{cond}} = \mathbb{E}[\mathbb{V}(\psi(Y_i, X_i, \theta_{\text{pop}})|X_i)]$.

Proof: The first result follows from Assumption 3 and Theorem 7.2 in Newey and McFadden (1994).

To prove the second result, we will first establish the joint asymptotic distribution of $\sqrt{N}(\hat{\theta} - \theta_{\text{pop}})$ and $\sqrt{N}(\theta_{\text{cond}}(\mathbf{X}) - \theta_{\text{pop}})$, and then we use this result to derive the asymptotic distribution of $\sqrt{N}(\hat{\theta} - \theta_{\text{cond}}(\mathbf{X}))$.

Let $h_0(\theta) = \mathbb{E}[\rho(X_i, \theta)]$. Because $h_0(\theta) = g_0(\theta)$, Assumption C.1*(ii)* also holds with $h_0(\theta)$ replacing $g_0(\theta)$. By Assumption C.1*(iv)* and the same argument as in the proof of Theorem 3.2, we obtain $\mathbb{E}\|\rho(X_i, \theta_{\text{pop}})\|^2 < \infty$.

Next, we will show that Assumption C.1*(v)* also holds with $\hat{h}_N$ replacing $\hat{g}_N$ and h_0 replacing g_0. Notice that

$$\begin{aligned}
&\left\| \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \rho(X_i, \theta) - \rho(X_i, \theta_{\text{pop}}) - \mathbb{E}\big[\rho(X_i, \theta)\big] \right\| \\
&\qquad = \left\| \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \mathbb{E}\Big[\psi(Y_i, X_i, \theta) - \psi(Y_i, X_i, \theta_{\text{pop}}) - \mathbb{E}\big[\psi(Y_i, X_i, \theta)\big] \;\Big|\; X_i \Big] \right\| \\
&\qquad \leq \mathbb{E}\left[\left\| \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \big(\psi(Y_i, X_i, \theta) - \psi(Y_i, X_i, \theta_{\text{pop}}) - \mathbb{E}\big[\psi(Y_i, X_i, \theta)\big] \big) \right\| \;\Big|\; \mathbf{X} \right].
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\sup_{\|\theta - \theta_{\text{pop}}\| \leq \delta_N} \left\| \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \rho(X_i, \theta) - \rho(X_i, \theta_{\text{pop}}) - \mathbb{E}\big[\rho(X_i, \theta)\big] \right\| \\
&\qquad \leq \mathbb{E}\left[\sup_{\|\theta - \theta_{\text{pop}}\| \leq \delta_N} \left\| \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \big(\psi(Y_i, X_i, \theta) - \psi(Y_i, X_i, \theta_{\text{pop}}) - \mathbb{E}\big[\psi(Y_i, X_i, \theta)\big] \big) \right\| \;\Big|\; \mathbf{X} \right].
\end{aligned}$$

By Markov's inequality, to show that the right hand side of last equation converges to zero in probability, it is enough to show that its expectation converges to zero:

$$\mathbb{E}\left[\sup_{\|\theta - \theta_{\text{pop}}\| \leq \delta_N} \left\| \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \big(\psi(Y_i, X_i, \theta) - \psi(Y_i, X_i, \theta_{\text{pop}}) - \mathbb{E}\big[\psi(Y_i, X_i, \theta)\big] \big) \right\| \right] \to 0.$$

Last equation holds by Lemma 2.3.11 in Van Der Vaart and Wellner (1996). The rest of the proof is as for Theorem 3.2.

C.3 Application to quantile regression

Let $F_{Y|X=x}(y) = \Pr(Y_i \leq y | X_i = x)$. For quantile regression:

$$\psi(Y_i, X_i, \theta) = X_i(I_{[Y_i - X_i'\theta \leq 0]} - \tau),$$

and

$$\rho(X_i, \theta) = X_i(F_{Y|X=X_i}(X_i'\theta) - \tau)$$

for some $\tau \in (0,1)$.

Assume that $\mathbb{E}\|X_i\| < \infty$ and that $I_{[Y_i - X_i'\theta \leq 0]}$ is continuous at each $\theta \in \Theta$ with probability one. Define θ_{pop} such that $\mathbb{E}[\psi(Y_i, X_i, \theta_{\text{pop}})] = 0$ and assume that this equation has a unique solution. Then, Theorem 1 implies: $\sqrt{N}(\hat{\theta} - \theta_{\text{pop}}) \xrightarrow{p} 0$ and $\sqrt{N}(\hat{\theta} - \theta(\mathbf{X})) \xrightarrow{p} 0$.

Next we verify the Assumptions of Theorem C.1. For the quantile regression estimator, it can be shown:

$$\left\| \frac{1}{N} \sum_{i=1}^{N} X_i(I_{[Y_i \leq X_i'\hat{\theta}]} - \tau) \right\| = o_p(1/\sqrt{N})$$

(see Powell, 1984). Notice that $g_0(\theta) = \mathbb{E}[X_i(F_{Y|X=X_i}(X_i'\theta) - \tau)]$. Let $\mathcal{B}$ be an open neighborhood of θ_{pop}. Assume that for almost all x in the support of X_i the function $F_{Y|X=x}(y)$ is continuously differentiable for all $y = x'\theta$ such that $\theta \in \mathcal{B}$, with bounded derivative $f_{Y|X=x}(x'\theta)$. Now if $\mathbb{E}[\|X_i\|^2] < \infty$, this implies (see, e.g., Lemma 3.6 in Newey and McFadden (1994)):

$$\Gamma = \mathbb{E}[f_{Y|X=X_i}(X_i'\theta_{\text{pop}}) X_i X_i'].$$

Assume that Γ is non-singular. $\mathbb{E}[\|X_i\|^2] < \infty$ also implies $\mathbb{E}\left[\|\psi(Y_i, X_i, \theta_{\text{pop}})\|^2\right] < \infty$. Assumption 3*(v)* is left to be verified. First, notice that each component of $\psi(y, x, \theta)$ is Euclidean for the envelope $\max\{\|x\|, 1\}$, as defined in Pakes and Pollard (1987). Because $\mathbb{E}[\|X_i\|^2] < \infty$, this envelope is square-integrable. Because $I_{[Y_i - X_i'\theta \leq 0]}$ is continuous with probability one at $\theta = \theta_{\text{pop}}$, Lemma 2.17 in Pakes and Pollard (1987) implies:

$$\sup_{\|\theta - \theta_{\text{pop}}\| \leq \delta_N} \sqrt{N}\|\hat{g}_N(\theta) - \hat{g}_N(\theta_{\text{pop}}) - g_0(\theta)\| \xrightarrow{p} 0.$$

As a result, we obtain that Theorem 3.2 holds with

$$\Delta_{\text{pop}} = \mathbb{E}[X_i(I_{[Y_i - X_i'\theta_{\text{pop}} \leq 0]} - \tau)^2 X_i'],$$

and

$$\Delta_{\text{cond}} = \mathbb{E}[X_i \mathbb{V}(I_{[Y_i - X_i'\theta \leq 0]} | X_i) X_i'].$$

Under correct specification, $\mathbb{E}[I_{[Y_i - X_i'\theta \leq 0]} | X_i] = \tau$, so $\mathbb{V}(I_{[Y_i - X_i'\theta \leq 0]} | X_i) = \mathbb{E}[(I_{[Y_i - X_i'\theta \leq 0]} - \tau)^2 | X_i] = \tau(1-\tau)$ and $\Delta_{\text{cond}} = \Delta_{\text{pop}}$.